An Analysis for some Methods and Algorithms of Quantum Chemistry

---

vorgelegt von
Dipl. Math. **Thorsten Rohwedder**
aus Preetz, Holstein

Von der Fakultät II - Mathematik und Naturwissenschaften -
der Technischen Universität Berlin
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften Dr. rer. nat.
genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. rer. nat. Martin Skutella (TU Berlin)
Berichter/Gutachter:	Prof. Dr. rer. nat. Christian Lubich (Univ. Tübingen)
	Prof. Dr. rer. nat. Reinhold Schneider (TU Berlin)
	Prof. Dr. rer. nat. Harry Yserentant (TU Berlin)

Tag der wissenschaftlichen Aussprache: 15.11. 2010

Berlin 2010
D 83

Dedicated to the people and things
without whom this work would not have been possible:

To my parents,
without whom I would not be where I am now,

to the people who have guided my way through science,
in particular to Reinhold Schneider,
Alexander Auer and Etienne Emmrich,

to the Universities of Kiel and Berlin
for providing the necessary financial support,

to all the friends who have accompanied me on the way,
and last, but not least, to Rock'n'Roll.

Preface and overview

More than 80 years after Paul Dirac stated that "the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are [...] completely known"[62], the development of "approximate practical methods of applying quantum mechanics", demanded by Dirac in the same breath, is still a highly active field of research at the crossroads of physics, chemistry, applied mathematics and computer science.

This circumstance is mainly owed to the interplay of two facts: On the one hand, the development of modern day computers has seen a phase of almost exponential growth at the end of the last century, so that calculations of theoretical chemistry and molecular physics have become competitive with practical experiments or at least often allow useful predictions of empirical parameters that can assist practical investigations. On the other hand, the solution of equations formulated in quantum mechanics is an exceedingly high-dimensional and thus computationally demanding problem, while at the same time, an extremely high accuracy is needed in order to obtain results utilizable in practice. Even nowadays, small to medium-sized quantum chemical problems push the limits of commonly available computational resources. To efficiently treat the variety of practical problems covered by the formalism of quantum mechanics, it is therefore indispensable to design highly problem-adapted methods and algorithms that balance the available computational resources against the respective required accuracy. These prerequisites have lead to a "zoo" of extremely sophisticated and well-developed methods and algorithms commonly used in quantum chemistry. Partly, the respective approaches are *ab initio*, i.e. the working equations are derived directly from the Schrödinger equation, as is for instance the case for the various variants of the Hartree-Fock method over perturbational methods, the Configuration Interaction (CI) and Coupled Cluster (CC) method and the recently revived CEPA method to reduced density matrix methods, to mention but the probably most important ones; to another part, they also integrate empirical parameters, as for instance in the successful Kohn-Sham model of density functional theory and the stochastic methods of Quantum Monte Carlo techniques do.[1]

Although the development of formal quantum mechanics and that of functional analysis are deeply interwoven, and although the theoretical properties of the Schrödinger equation and the Hamiltonian are quite well understood from a mathematical point of view (see Section 1), most of the practically relevant computational schemes mentioned above were introduced by physicists or chemists, and the actual algorithmic treatment of the electronic Schrödinger equation does only recently seem to have aroused the broader attention of the mathematical community. Therefore, although there have been various efforts in

[1]For an introduction and references to the respective methods, see e.g. [103, 201] for Hartree-Fock, [201, 142] for perturbational approaches, Section 2.1 of this work for references for the CI method and density functional theory, Section 3 for the Coupled Cluster method, [133, 208] for the CEPA method, [148] for the reduced density matrix methods and [78, 144] for a review of Quantum Monte Carlo techniques.

understanding the methods of quantum chemistry from a mathematical point of view[2] and to approach general problems in the numerical treatment of the electronic Schrödinger equation by means of concepts from mathematics,[3] the stock of available mathematically rigorous analysis of the present practically relevant methods of quantum mechanics and of the convergence behaviour of the algorithms used for their treatment is on the whole still relatively scarce. It will be subject of the present work to approach this shortcoming, that is, to provide a numerical analysis for certain aspects of some well-known methods of quantum chemistry.

The work is organized in four parts. The first part (Section 1) is an attempt to connect the world of mathematical physics to that of computational chemistry: Starting from the necessities imposed by the postulates of quantum mechanics, we introduce the operators and spaces needed to embed the main task of electronic structure calculation, i.e. the calculation of electronic states and energies, into a sound mathematical background; we review known theoretical results, prove some results needed later and derive the (Galerkin) framework that is in a wider sense the basis to all methods used in practical calculations in quantum chemistry.

From Section 2 onwards, we turn towards the actual algorithmic treatment of the equations derived in Section 1: Section 2, parts of which have already been published in [191], first provides a short introduction to the methods of Hartree-Fock, Kohn-Sham and CI; we then give a convergence analysis for a preconditioned steepest descent algorithm under orthogonality constraints, taylor-made for and commonly used in the context of Hartree-Fock and density functional theory calculations, but also providing a sensible algorithm for implementation of the CI method. Section 3, featuring some of the main achievements of this work, is dedicated to lifting the Coupled Cluster method, usually formulated in a finite dimensional, discretised subspace of a suitable Sobolev space H^1, to the continuous space H^1, resulting in what we will call the continuous Coupled Cluster method. To define the continuous method, some formal problems have to be overcome; afterwards, the results for the continuous methods will be used to derive existence and (local) uniqueness statements for discretisations and to establish goal-oriented *a-posteriori* error estimators for the Coupled Cluster method. The last part of this work (Section 4) features an analysis for the acceleration technique DIIS that is commonly used in quantum chemistry codes. To derive some (positive as well as negative) convergence results for DIIS, we establish connections to the well-known GMRES solver for linear systems as well as to quasi-Newton methods.

At the beginning of each of the sections, a more thorough introduction to their respective subject is given. Also, in Sections 2 – 4, the main results of the respective section will explicitly be referenced there. Mathematical objects and notions used in this work are either introduced explicitly, or the reader is referred to according literature. Please also note that the most frequently used notations are compiled in the list of symbols at the end of this work.

[2] Cf. e.g. [7, 42, 43, 186, 190] for recent works and also [15, 139, 140, 143, 153] for the properties of the Hartree-Fock method, already analysed to some extent in the 1970-80's.

[3] See e.g. [45, 79, 92, 214].

Preface and overview

Contents

1 A mathematical framework for electronic structure calculation

Since the hour of birth of formal quantum mechanics in the mid-1920s, it is known that the behaviour of non-relativistic [173] atoms and molecules can quite accurately be described by the (time-dependent) *Schrödinger equation* [192],

$$i\hbar\frac{d}{dt}\Psi \;=\; H_{\mathrm{mol}}\Psi. \tag{1.1}$$

This equation fixes the behaviour of a system consisting of a given number N of electrons and a prescribed number M nucleons of charges Z_k, $k \in M] := \{1, \dots, M\}$, exposed to a given outer potential V, by an accordingly constructed molecular Hamiltonian operator H_{mol}.
The solutions Ψ of the Schrödinger equation are so-called *wave functions* or *states*, depending on the coordinates $x_i \in \mathbb{R}^3, i \in N]$ of the N electrons, the coordinates $y_j \in \mathbb{R}^3, j \in M]$ of the M nuclei, a respective spin variable $s_i, s'_j \in \{\pm\frac{1}{2}\}$ for each of the particles, and a time coordinate $t \in \mathbb{R}$. For any fixed time $t \in \mathbb{R}$, a solution $\Psi(\cdot, t)$ of (1.1) is an element of the vector space

$$\mathcal{L}^2_{N,M} \;:=\; L^2(\mathbb{R}^{3(N+M)} \times \Sigma_{\mathrm{mol}}),$$

in which Σ_{mol} denotes a suitable discrete space modelling the spin variable. Of supreme interest to quantum chemistry are the stationary *bound states* of a given configuration of particles, which can be computed by solving the operator eigenvalue equation for H_{mol} [183]. Stationary solutions of the original equation (1.1) are then given by the eigenfunctions $\Psi \in \mathcal{L}^2_{N,M}$ of H_{mol}, multiplied by a phase factor determining the time dependence; the corresponding eigenvalue gives the total energy of the state.
In a next step, the eigenvalue problem for H_{mol} is usually reduced further to an *eigenvalue problem for an electronic Hamiltonian* H: The mass of the nuclei is more than 10^3 times greater than that of electrons, and this fact is used to justify [97, 117] the so-called *Born-Oppenheimer approximation* [32], approximating the quantum mechanical properties of a given configuration by computing only an *electronic* wave function,

$$\Psi\big((x_1, s_1), \dots, (x_N, s_N)\big) \;\in\; \mathcal{L}^2_N \;:=\; L^2\big(\mathbb{R}^{3N} \times \{\pm\frac{1}{2}\}^N\big).$$

Ψ now solely describes the behaviour of the electrons with coordinates $x_i \in \mathbb{R}^3$ and spins $s_i \in \{\pm\frac{1}{2}\},\ i \in N]$, while the M nuclei are now represented by point charges clamped at fixed positions $R_1, \dots, R_M$ and induce an outer field incorporated in the potential V. The benefits of this are that the space $\mathcal{L}^2_{N,M}$ is in this way replaced by the somewhat smaller (but in sensible discretisations unfortunately still extremely high-dimensional) space $\mathcal{L}^2_N$, and that the wave function now describes a set of *identical* particles.

In a first rough version, we may now phrase the main task of electronic structure calculation as follows: For a given configuration of nuclei, fixed at positions $R_1, \dots, R_M \in \mathbb{R}^3$ and carrying positive charges $Z_1, \dots, Z_M$, and for a given number N of electrons, calculate the possible bound states $\Psi \in \mathcal{L}^2_N$ and the according binding energies of this configuration. Essentially, it is due to John von Neumann [155] and Tosio Kato [113] that this rather informal formulation can be rephrased mathematically precisely in terms of self-adjoint unbounded operators on Hilbert spaces.[4] In the present first section of this work, we will take this mathematical framework as a starting point to develop a setting that combines the variational framework commonly used in numerical mathematics with the Second Quantization formalism that is often used in the context of quantum chemistry, thus equipping ourselves with a sensible background for a numerical analysis of the methods and algorithms of quantum chemistry. As well, we will supply many of the auxiliary means and notations needed in this work. For the sake of brevity, we will comprise only the specific framework needed for the description of electronic wave functions, that is, for a quantum mechanical system of N indistinguishable fermions.

In the Sections 1.1 and 1.2, the spaces and operators setting the general framework are introduced. At the end of Section 1.3, we will arrive at a weak formulation of the electronic Schrödinger equation which, due to an antisymmetry constraint and various invariances of the Hamiltonian H, can be decomposed to single computations on antisymmetric subspaces $\mathbb{L}^2_k$ of $\mathcal{L}^2$, belonging to a fixed z-spin number k. Note in this context that, presumably for notational convenience, the z-spin variable is neglected in most theoretical investigations of the electronic Schrödinger equation. Nevertheless, restrictions imposed by a fixed z-spin reduce the size of the underlying tensor basis and thus the computational complexity, and certain spin selection schemes are therefore integrated in almost every quantum chemical code. Recalling the aim of this section, namely to embed the methods used in quantum chemistry into a sound mathematical background, we have therefore decided to explicitly formulate the electronic Schrödinger equation in terms of the spin spaces $\mathbb{L}^2_k$, and apologize for the notational inconvenience aroused by this. In Section 1.4, we then prepare a Galerkin method for the weak Schrödinger equation by constructing tensor bases for the antisymmetric spaces $\mathbb{L}^2_k$. Section 1.5 will be dedicated to reformulation of the weak Schrödinger equation in terms of annihilation and creation operators borrowed from the formalism of *Second Quantization*, which for more sophisticated methods of Quantum Chemistry, as e.g. the Coupled Cluster method (Section 3), simplifies the derivation of implementable equations significantly. In this context, evaluation rules for the matrix elements of H with respect to the constructed tensor basis will be given, thus equipping us with the necessary means for a Galerkin discretisation of the electronic Schrödinger equation. Finally, we compile some general results that will be needed later, and close in Section 1.7 with a discussion of topics related to the discretisation of the electronic Schrödinger equation.

[4] For a more thorough history of quantum mechanics, see the timeline in [111] or the more textbook-like [169].

For further reading on the subjects of this section, we refer to [64, 105, 177, 179, 206, 207] for the functional analytic background, to [28, 175, 202] for a general introduction to mathematical formalism of quantum mechanics, to the monographs [105, 178, 179] and to the reviews [111, 197, 198] for an overview of results on Hamiltonians of quantum mechanics and their eigenfunctions, and to the monographs [43, 48, 214] for the concrete application of the electronic Schrödinger equation. The treatment given here is based on the axioms of nonrelativistic quantum physics [8, 28, 152, 175]; in particular, relativistic effects are excluded throughout this work.

1.1 General setting

In this first section, we will introduce the (tensor product) Lebesgue and Sobolev spaces used in electronic structure calculation, as well as operators acting on them. In particular, we will define the Hamiltonian H of an N-electron system.

(i) The Lebesgue space $\mathcal{L}^2$ for N electrons. As usual, $L^2(\Omega) = L^2(\Omega, \mathbb{C})$ will in this work denote the space of complex-valued, measurable, square-integrable functions defined on a measure space Ω [20]. In the formalism of quantum mechanics, a single electron is described by a normed *state function*

$$\chi(x, s) \ \in \ L^2(\mathbb{R}^3 \times \Sigma),$$

depending on a spatial variable $x \in \mathbb{R}^3$ and a spin variable $s \in \Sigma = \{+\frac{1}{2}, -\frac{1}{2}\}$. In a transition that is mainly motivated by gas statistics [175], a system consisting of N electrons is represented by a normed[5] element Ψ from the according N-fold tensor product space[6]

$$\mathcal{L}^2 \ := \ \mathcal{L}^2_N \ := \ \bigotimes_{i=1}^{N} L^2(\mathbb{R}^3 \times \Sigma). \tag{1.2}$$

A quantum mechanical entity $\Psi \in \mathcal{L}^2$ describing a system of N electrons is thus a function depending on N spatial variables $x_1, \ldots, x_N \in \mathbb{R}^3$, which we will in the following also collectively denote by a vector $X = (x_1, \ldots, x_N)$, and of N respective spin variables, $s_1, \ldots, s_N \in \{-\frac{1}{2}, \frac{1}{2}\}$, compiled in one spin vector

$$\sigma \ = \ (s_1, \ldots, s_N) \ \in \ \Sigma^N \ := \ \left\{-\frac{1}{2}, \ \frac{1}{2}\right\}^N, \tag{1.3}$$

[5] Note that the norm condition is in accordance with a probabilistic interpretation of the wave function: Integration of Ψ over a set of volumes $\Omega_1, \ldots, \Omega_N \subseteq \mathbb{R}^3$ and respective spins $s_1, \ldots, s_N \in \{\pm\frac{1}{2}\}$ will give the probability of simultaneously finding one electron in Ω_1 with spin s_1, one in Ω_2 with spin s_2 and so on.

[6] We suppress the suffix N here, $\mathcal{L}^2 := \mathcal{L}^2_N$, to keep notations short. In the following, this will often be done in case the number N of electrons under consideration is fixed.

so that we obtain the more compact notation $\Psi = \Psi(X, \sigma)$.

Because the tensor product space (1.2) sets the general framework for the description of N-electron systems, let us shortly recall some properties of abstract tensor spaces $\otimes_{i=1}^N V$ formed from a set of N identical Hilbert spaces $V = V_1 = \ldots = V_N$ with inner product $\langle \cdot, \cdot \rangle_V$, see [94, 184, 210] for more thorough introductions to the subject. To obtain these tensor product spaces and an inner product on them, one builds in a first step from the N copies of V the *algebraic tensor space* $(\otimes_{i=1}^N V)_{\text{alg}}$. On this space, an inner product can be obtained by at first defining for *elementary tensors*

$$v \;=\; \otimes_{i=1}^N v_i \;:=\; v_1 \otimes \ldots \otimes v_N, \qquad w = \otimes_{i=1}^N w_i \tag{1.4}$$

that

$$\langle v, w \rangle_\otimes \;:=\; \langle \otimes_{i=1}^N v_i,\;\; \otimes_{i=1}^N w_i \rangle_\otimes \;\;:=\;\; \prod_{i=1}^N \langle v_i, w_i \rangle_V, \tag{1.5}$$

and by then (bi-)linearly extending this definition to all of $(V_1 \otimes \ldots \otimes V_N)_{\text{alg}}$. The (analytic) *tensor product space* $\otimes_{i=1}^N V$ is then obtained by taking the closure of $(\otimes_{i=1}^N V)_{\text{alg}}$ with respect to the norm $\| \cdot \|_\otimes$ induced by this inner product,

$$\otimes_{i=1}^N V := \overline{(\otimes_{i=1}^N V)_{\text{alg}}}^{\|\cdot\|_\otimes}.$$

If $B_V = (v^{(k)})_{k \in I}$ is a basis of V, a basis for $\otimes_{i=1}^N V$ is given by

$$\mathcal{B} = \{ v^{(k_1)} \otimes \ldots \otimes v^{(k_N)} | k_1 \ldots, k_N \in I \}; \tag{1.6}$$

it is orthonormal if and only if B_V is orthonormal. Furthermore, if $\dim V$ is finite,

$$\dim\, (\otimes_{i=1}^N V) = (\dim V)^N. \tag{1.7}$$

From the numerical point of view, equation (1.7) displays an unfortunate fact of quantum chemistry (and of tensor product spaces in general): The dimensions of the spaces under consideration depend exponentially on the number N of particles. Thus, they are for any but very small N extremely high dimensional, a fact that is (using a phrase lent from [25]) sometimes termed the "curse of dimensionality".

For the space $\mathcal{L}^2$ constructed in the way outlined above, there holds

$$\mathcal{L}^2 = L^2(\mathbb{R}^{3N} \times \Sigma^N) \tag{1.8}$$

by application of the Fubini-Tonelli theorem (see e.g. [20]) to (1.5). Thus, the inner product on $\mathcal{L}^2$ is given by

$$\langle \Psi, \Psi' \rangle \;\; := \langle \Psi, \Psi' \rangle_{\mathcal{L}^2} \;\; := \;\; \sum_{\sigma \in \Sigma^N} \int\limits_{\mathbb{R}^{3N}} \Psi(X, \sigma) \overline{\Psi'(X, \sigma)}\, dX. \tag{1.9}$$

Note though that if Ψ and Ψ' can be represented in a specified tensor basis (1.6), the inner product on $\mathcal{L}^2$ can be broken down into the inner products amongst the basis functions, which may be computed according to (1.5) - a fact that will be useful later on. The induced norm on $\mathcal{L}^2$ will in the following be denoted by

$$\|\Psi\|^2 \quad := \|\Psi\|^2_{\mathcal{L}^2} \quad := \quad \sum_{\sigma\in\Sigma^N} \int\limits_{\mathbb{R}^{3N}} |\Psi(X,\sigma)|^2 \, dX; \tag{1.10}$$

the normalization condition for a state function is therefore

$$\|\Psi\|^2 = 1. \tag{1.11}$$

(ii) Observables on $\mathcal{L}^2$. On the state space $\mathcal{L}^2$, physical quantities (or *observables*) like energy, spin, angular momentum etc. of a quantum mechanical system are (in contrast to classical physics) represented by self adjoint operators $\mathcal{O} : \mathcal{D}(\mathcal{O}) \to \mathcal{L}^2$, where $\mathcal{D}(\mathcal{O})$ is a dense subset of $\mathcal{L}^2$. The outcome of a the measurement of the observable $\mathcal{O}$ imposed on a state $\Psi \in \mathcal{L}^2$ is not deterministic, but of statistic nature: If

$$\mathcal{O} = \int_{\mathbb{R}} \lambda \, dE(\lambda) \tag{1.12}$$

is the spectral decomposition [64] of $\mathcal{O}$, the probability of measuring a value $m \in]a,b]$ for $\mathcal{O}$ is given by

$$P(m \in]a,b]) = \langle \Psi, (E(b) - E(a))\Psi \rangle.$$

Therefore, the *spectral properties* of observables, especially that of the Hamiltonian H of the system, measuring its total energy, are of primary importance in quantum mechanics and quantum chemistry. Note that the self-adjointness of observables implies that the spectrum of an observable is real, so that a measurement produces a real value, like one would expect from quantities measured in classical physics.
A particularly simple class of N-particle observables $\mathcal{O}$ acting upon an N-particle wave function is constituted by those that measure the sum of observables for the single particles; for example, the total kinetic energy of a system of N particles is given by the sum of the kinetic energies of the single particles. The corresponding mathematical construction uses the following definition.

Definition 1.1. *(Kronecker products of operators)*
Let X be a Hilbert space and $A_i, i \in N]$ a set of N densely defined symmetric operators, $A_i : X \supseteq \mathcal{D}(A_i) \to X$. The *tensor product* or *Kronecker product* $A_1 \otimes \ldots \otimes A_N$ of those operators is defined by

$$\begin{aligned} A_1 \otimes \ldots \otimes A_N : \quad D(A_1 \otimes \ldots \otimes A_N) \quad &:= \quad \mathcal{D}(A_1) \otimes \ldots \otimes \mathcal{D}(A_N) \to \otimes_{i=1}^N X, \\ (A_1 \otimes \ldots \otimes A_N)(f_1 \otimes \ldots \otimes f_N) \quad &:= \quad A_1 f_1 \otimes \ldots \otimes A_N f_N \end{aligned} \tag{1.13}$$

for elementary tensors, and then continuation by linear extension and taking the closure with respect to graph norm [206] induced by the tensor product norm on $\otimes_{i=1}^{N} X$. In particular, we will often encounter the "lifting" of an operator $A : X \supseteq \mathcal{D}(A) \to X$ to an operator A_N on $\otimes_{i=1}^{N} X$ by

$$\begin{aligned} A_N &:= \Big(A \otimes I \otimes \ldots \otimes I \;+\; I \otimes A \otimes \ldots \otimes I \;+\; \ldots \;+\; I \otimes \ldots \otimes I \otimes A \Big) \\ &= \sum_{i=1}^{N} \Big(\otimes_{k=1}^{N} \left(\delta_{k,i} A + (1 - \delta_{k,i}) I\right) \Big) \;=:\; \sum_{i=1}^{N} \hat{A}_i. \end{aligned} \tag{1.14}$$

If the context is clear, the suffix N will often be dropped, i.e. A_N will also simply be denoted by A.

Examples for operators built from sums of Kronecker products will be the spin operator introduced in Section 1.2 and the Hamiltonian to be defined in part (iv) of this section.

(iii) Sobolev spaces. The Hamiltonian of a quantum mechanical system contains differential operators and can therefore not be defined on all of $\mathcal{L}^2$, but only on the Sobolev space $\mathcal{H}^2 \subseteq \mathcal{L}^2$. We give the more global definition of Sobolev spaces that will be used in various contexts later.

Definition 1.2. *(Sobolev spaces $H^t(\Omega)$)*
Let Ω be a measure space, and for a function $u(x) \in L^2(\Omega)$, let $\mathcal{F}u(\omega) \in L^2(\Omega)$ denote its Fourier transform [182]. On the subspace

$$C_0^\infty(\Omega) \;\subseteq\; L^2(\Omega)$$

of infinitely often differentiable functions with compact support, we define for any real $t \geq 0$ the inner product

$$\langle u, v \rangle_t \quad := \quad \langle (1 + |\omega|^2)^t \; \mathcal{F}u(\omega),\; \overline{\mathcal{F}v(\omega)} \; \rangle. \tag{1.15}$$

The Sobolev space $H^t(\Omega)$ is the subspace of $\mathcal{L}^2$ obtained by closing C_0^∞ with respect to the norm $\| \cdot \|_t$ induced by $\langle \cdot, \cdot \rangle_t$. In particular, we will denote by

$$\mathcal{H}^t \quad := \quad \mathcal{H}_N^t \quad := \quad H^t(\mathbb{R}^{3N} \times \Sigma^N) \tag{1.16}$$

the Sobolev subspaces of $\mathcal{L}^2$.
On $H^t(\Omega)$, the canonical norm is given by $\| \cdot \|_t$ and will be denoted this way throughout this work. The dual space of $H^t(\Omega)$ will be denoted by $H^{-t}(\Omega)$.

□

We will in this work mostly be concerned with suitable subspaces of the Sobolev spaces $H^1(\Omega)$ and $H^2(\Omega)$. For any $t \geq 0$, $H^t(\Omega)$ is a Hilbert space with the above inner product (1.16) and dense in $L^2(\Omega)$, see [182] for details. Note also that for $t > 0$, $\mathcal{H}^t$ is not equal to the tensor product space $\mathcal{H}^t_\otimes := \otimes_{i=1}^N H^t(\mathbb{R}^3 \times \Sigma)$ constructed from the Hilbert spaces $H^t(\mathbb{R}^3 \times \Sigma)$ according to the proceeding outlined in part (i) of this section: Due to the mixed product terms arising in the inner products (1.5), additional conditions are imposed on the mixed derivatives of functions from $\mathcal{H}^t_\otimes$; therefore, $\mathcal{H}^t_\otimes \subset \mathcal{H}^t$.

(iv) The electronic Hamiltonian. The electronic Hamilton operator H of a system of N electrons, defined on $\mathcal{H}^2$, is the observable measuring the (nonrelativistic) total energy of a system of N electrons exposed to an outer potential. In particular, the spectrum of H determines the energy values that electronic configurations under description can attain. These values are not only of interest for itself, determining e.g. bonding, ionization and reaction energies for molecules; also, one can derive several other physical and chemical quantities like equilibrium geometries, bond lengths, vibrational frequencies, energy gradients and other molecular properties by geometry optimization or by deriving the energy with respect to certain parameters, see e.g. [103, 201].
We will now introduce the Hamiltonian H for a purely electronic system, exposed to a field induced by a fixed configuration of nuclei. H is obtained from the classical expression for the energy of a system [84] by the so-called correspondence principle [175]. The formulation will be given in atomic units [196], so that no constants unnecessary for the mathematical treatment are involved in the Schrödinger equation; consequently, energies will be measured in *Hartree*.[7] Note that the below choices for the kinetic and potential energy operators T and V defined in this context are of axiomatic nature, not adequate any more in relativistic quantum mechanics. Also, in order to obtain the energy of the whole molecule (in terms of the Born-Oppenheimer approximation), one additionally has to add the term $R := \sum_{k=1}^M \sum_{\ell=1,\ell\neq k}^M Z_k Z_\ell / |R_k - R_\ell|$, describing the (classical) interaction between the nuclei, to the electronic Hamiltonian introduced below. Because R only adds a constant shift to H, it is in practice usually precalculated and added afterwards. The following definition for the electronic Hamiltonian H presumes (in connection with the definition of observables as self-adjoint operators) that H is well defined and self-adjoint on the Sobolev space $\mathcal{H}^2$. That this indeed holds follows from Rellich's theorem [178] and Kato [113], see [175] for a compilation of both results.[8,9]

[7] 1 Hartree $= 1E_h = 4.35974417(75) \cdot 10^{-18} J$

[8] Strictly speaking, the cited results only show that the position space Hamiltonian H_X defined below is self-adjoint. It is not hard to see though that this is equivalent to the self-adjointness of H.

[9] For the self-adjointness of related molecular Hamiltonians using different potentials, also cf. [28, 105, 178, 197, 206] and the references given in the latter.

Definition 1.3. (The electronic Hamiltonian H)

The nonrelativistic electronic Hamiltonian $H : \mathcal{H}^2 \to \mathcal{L}^2$ is defined via a position space Hamiltonian

$$H_X \;:\; H^2(\mathbb{R}^{3N}) \;\to\; L^2(\mathbb{R}^{3N})$$

acting on $\Psi(X,\sigma) \in \mathcal{D}(H)$ spin-component-wise, i.e. with $\Sigma = \{\sigma_1, \ldots, \sigma_{2^N}\}$,

$$H\Psi(X,\sigma) \;=\; H \begin{pmatrix} \Psi(X,\sigma_1) \\ \Psi(X,\sigma_2) \\ \vdots \\ \Psi(X,\sigma_{2^N}) \end{pmatrix} \;:=\; \begin{pmatrix} H_X\Psi(X,\sigma_1) \\ H_X\Psi(X,\sigma_2) \\ \vdots \\ H_X\Psi(X,\sigma_{2^N}) \end{pmatrix}. \tag{1.17}$$

The position space Hamiltonian H_X used here is defined as the sum of the observables measuring kinetic and potential energy, $H_X = T + V$. To define $T : H^2(\mathbb{R}^{3N}) \to L^2(\mathbb{R}^{3N})$, we extend the scaled 3-dimensional Laplacian,[10]

$$-\frac{1}{2}\Delta : \quad H^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3), \quad \varphi \;\mapsto\; -\frac{1}{2}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\varphi, \tag{1.18}$$

according to Definition 1.1 to the tensor product space $\mathcal{H}^2$,

$$T \;=\; T_N \;:=\; -\frac{1}{2}\sum_{i=1}^{N} \hat{\Delta}_i. \tag{1.19}$$

The potential energy[11] *observable V is given by a multiplication operator,*

$$V :\; \Phi(X) \;\mapsto\; V(X)\cdot\Phi(X), \qquad V(X) \;=\; \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{1}{|x_i - x_j|} \;-\; \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{Z_k}{|x_i - R_k|}.$$

Thus,

$$H_X \;=\; \underbrace{-\frac{1}{2}\sum_{i=1}^{N}\hat{\Delta}_i}_{=:T} \;+\; \underbrace{\underbrace{\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{1}{|x_i - x_j|}}_{=:V_{el}} \;-\; \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{|x_i - R_k|}}_{=:V_{core}}}_{=:V}. \tag{1.20}$$

[10] This choice can be motivated heuristically by de Broglie's law, see the treatment in [189], or more strictly by replacing in the classical relation $E_{kin} = \frac{p^2}{2m}$ between kinetic energy E_{kin}, momentum p and mass m the momentum variable by the associated quantum mechanical observable according to the correspondence principle of quantum mechanics, see [175].

[11] For the potential energy part V, it is custom to choose a potential which reflects all the inner and outer forces acting upon the system. In the electronic Hamiltonian, the first term of V models the repulsive Coulomb interaction amongst the electrons, while the second reflects the attractive electron-nucleon forces (described within the Born-Oppenheimer approximation as semi-classical interaction between quantum-mechanical electrons and point-like nuclei.)

1.2 The Pauli principle and invariant subspaces of the Hamiltonian

In this section, we will introduce the Pauli principle that enforces admissible wave functions to be antisymmetric, as well as some of the invariant subspaces of the Hamiltonian H that are nowadays the most common ones used to reduce computational complexity in practice. For a more thorough overview and an identification with "symmetries" imposed by other observables on $\mathcal{L}^2$, see [169].

(i) The Pauli principle and the N-fermion state space $\hat{\mathcal{L}}^2$. A restriction on admissible solutions of the electronic Schrödinger equation is imposed by the so-called (generalized) Pauli principle: If a system incorporates only identical particles, like in our case, a system of N electrons does, the particles cannot be distinguished from each other by any measurement, i.e. by the outcome of the action of a densely defined, self adjoint linear operator on the state space. From this postulate, it follows [214] that wave functions describing a system of N identical particles have to be either symmetric or antisymmetric with respect to exchange of particle coordinates. The (semi-empirical) spin-statistics relation [166, 73] identifies antisymmetric wave functions with multi-particle systems of *fermions*, i.e. particles with *half integer* spin. Electrons, like protons and neutrons, are such particles of half-integer spin, a fact experimentally supported for instance by a splitting of the hydrogen spectral lines called fine structure (see standard textbooks on physics, e.g. [84]). Therefore, the wave function Ψ of an N-electron system has to be completely antisymmetric with respect to permutations of the particle indices, meaning that it changes sign under each transposition of particle indices. Formulated more generally,

$$\Psi((x_1, s_1), \ldots, (x_N, s_N)) \quad = \quad \mathrm{sgn}(\pi) \cdot \Psi((x_{\pi(1)}, s_{\pi(1)}), \ldots, (x_{\pi(N)}, s_{\pi(N)})) \tag{1.21}$$

has to hold for all permutations π operating on the N indices of Ψ and for any point[12] $(X, \sigma) \in \mathbb{R}^{3N} \times \Sigma^N$. The set of admissible wave functions for a system of N identical fermions thus reduces to the subspace $\hat{\mathcal{L}}^2$ of $\mathcal{L}^2$ containing only the antisymmetric functions of $\mathcal{L}^2$. We will define this space more precisely now.

Definition 1.4. *(Antisymmetry projector)*
The *antisymmetry projector* $\mathcal{P}^a : \mathcal{L}^2 \to \mathcal{L}^2$ is defined by its action on arbitrary state functions $\Psi = \Psi((x_1, s_1), \ldots, (x_N, s_N)) \in \mathcal{L}^2$, given by

$$\mathcal{P}^a\Psi = \frac{1}{N!} \sum_{\pi \in S(N)} (-1)^{\mathrm{sgn}(\pi)} \Psi((x_{\pi(1)}, s_{\pi(1)}), \ldots, (x_{\pi(N)}, s_{\pi(N)})), \tag{1.22}$$

where the sum runs over the permutational group $S(N)$ on N elements, operating on the indices of Ψ.

[12] Although $\Psi \in \mathcal{L}^2$ is only determined up to null sets, we will see later (see Section 1.3(v)) that electronic wavefunctions Ψ are continuous, i.e. have a continuous representant; therefore, the equality indeed holds everywhere on $\mathbb{R}^{3N} \times \Sigma^N$ in this sense.

Lemma 1.5. *$\mathcal{P}^a$ is an $\mathcal{L}^2$-orthogonal projector, mapping onto a closed subspace of $\mathcal{L}^2$ containing the antisymmetric functions of $\mathcal{L}^2$. For any $t \geq 0$, it boundedly maps $\mathcal{H}^t \to \mathcal{H}^t$ with norm $\|\mathcal{P}^a\|_t = 1$.*

Proof. It is easy to verify that $\mathcal{P}^a$ is a linear projector on the tensor product space $\mathcal{L}^2$, and that it maps $\mathcal{L}^2$ to the antisymmetric functions by definition. Because for all permutations π acting on the indices of a wave function, $\|\pi\Psi\|_t = \|\Psi\|_t$ holds by definition of the inner product on $\mathcal{H}^t$, $\|\mathcal{P}^a\Psi\|_t \leq \|\Psi\|_t$ by the triangle inequality, so $\|\mathcal{P}^a\|_t = 1$ is obtained by mapping any antisymmetric $\hat{\Psi} \in \mathcal{H}^t$ with $\mathcal{P}^a$. It is not hard to show that $\mathcal{P}^a$ is symmetric with respect to the $\mathcal{L}^2$-inner product, so we skip the proof. In particular, because $\mathcal{P}^a$ is defined on all of $\mathcal{L}^2$, $\mathcal{P}^a$ is self-adjoint, and range($\mathcal{P}^a$) is closed because $\mathcal{P}^a$ is a projector.

□

Definition 1.6. *(Antisymmetric spaces $\hat{\mathcal{L}}^2, \hat{\mathcal{H}}^t$)*
We define the *space of antisymmetric N-electron functions*[13] as

$$\hat{\mathcal{L}}^2 \;:=\; \hat{\mathcal{L}}^2_N \;:=\; \wedge_{i=1}^N L^2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \;=\; \operatorname{range}\mathcal{P}^a \;=\; \mathcal{P}^a\mathcal{L}^2. \tag{1.23}$$

Also, for $t \geq 0$, we let

$$\hat{\mathcal{H}}^t \;:=\; \hat{\mathcal{H}}^t_N \;:=\; \mathcal{H}^t \cap \hat{\mathcal{L}}^2 \tag{1.24}$$

be the spaces of antisymmetric functions of Sobolev regularity t. Note that $\hat{\mathcal{H}}^t$ is closed with respect to the $\mathcal{H}^t$-norm due to Lemma 1.5.

□

[13] As before, we will drop the suffix N, e.g. $\hat{\mathcal{L}}^2 := \hat{\mathcal{L}}^2_N$ if the number of electrons under consideration is fixed. The notation $\wedge_{i=1}^N L^2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$, only used one time below, will not be used in the further work, but was taken up at this point because it is a common notation in other works on electronic structure calculation.

(ii) Spin symmetries: The spaces $\mathcal{L}^2_k$ of fixed z-spin. As an example for a one-particle operator defined on $\mathcal{L}^2$, we will now introduce the operator S^z_N measuring the spin of an N-electron system with respect to a preferential direction, commonly chosen along the z-axis. For a more thorough introduction to the matter of spin, see [136].

Definition 1.7. *(One-electron and N-electron z-spin operators)*
The *one-electron z-spin operator*

$$S^z : L^2(\mathbb{R}^3 \times \Sigma) \to L^2(\mathbb{R}^3 \times \Sigma)$$

acts solely on the spin variable of a one-electron wave function $\chi(x,s)$ having two spin components $\chi(x,1/2), \chi(x,-1/2) \in L^2(\mathbb{R}^3)$. S^z can therefore be defined in terms of one of the so-called Pauli matrices, namely

$$S^z = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad S^z\varphi(x,s) \;=\; \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} \varphi(x,\frac{1}{2}) \\ \varphi(x,-\frac{1}{2}) \end{pmatrix} \;=\; \begin{pmatrix} \frac{1}{2}\varphi(x,\frac{1}{2}) \\ -\frac{1}{2}\varphi(x,-\frac{1}{2}) \end{pmatrix}.$$

The *N-electron z-spin operator* S^z_N, measuring the total $z-$spin of the system, is now defined by using Definition 1.1 to set

$$S^z_N \;:=\; \sum_{i=1}^{N} \hat{S}^z_i : \quad \mathcal{L}^2 \to \mathcal{L}^2. \tag{1.25}$$

□

Obviously, the eigenvalues of the one-electron operator S^z are $\zeta_{1,2} = \pm\frac{1}{2}$, and the corresponding eigenfunctions are all nontrivial functions φ (in L^2-sense) for which $\varphi(x,-\frac{1}{2}) = 0$ or $\varphi(x,\frac{1}{2}) = 0$ respectively. S^z is a symmetric operator defined on all of $L^2(\mathbb{R}^3 \times \Sigma)$, so that S^z_N is a symmetric operator defined on all of $\mathcal{L}^2$, thus self-adjoint and therefore a quantum mechanical observable.

Definition 1.8. (Spin numbers and the spin spaces $\mathcal{L}^2_k, \mathcal{H}^t_k$)

We will call $k \in \{0,\dots,N\}$ a *spin number*. Let us abbreviate by

$$\mathrm{spin}(N) := \Big\{ -\frac{N}{2} + k \mid k \in \{0,\dots,N\}\Big\}$$

the eigenvalues of the z-spin operator S^z_N. For $\zeta_k \in \mathrm{spin}(N)$ and $t \geq 0$, we define the spin spaces

$$\mathcal{L}^2_k \;:=\; \mathrm{Eig}_{\zeta_k}(S^z_N) \;:=\; \{\Psi \in \mathcal{L}^2 \mid S^z_N\Psi = s_k\Psi\}, \quad \mathcal{H}^t_k \;:=\; \mathcal{L}^2_k \cap \mathcal{H}^t. \tag{1.26}$$

□

(iii) Invariant subspaces of H.

Definition/Lemma 1.9. *(Decomposition of H into invariant subspaces)*

(i) For the antisymmetrization projector $\mathcal{P}^a$, there holds $H\mathcal{P}^a\Psi = \mathcal{P}^a H\Psi$ for all $\Psi \in \mathcal{H}^2$. Thus, H maps $\hat{\mathcal{H}}^2 \to \hat{\mathcal{L}}^2$.

(ii) The spaces

$$\mathcal{L}^2_{\mathbb{R}} := \{\Psi \in \mathcal{L}^2 \mid Im\ \Psi = 0\}, \quad \mathcal{L}^2_{\mathbb{C}} := \{\Psi \in \mathcal{L}^2 \mid Re\ \Psi = 0\}$$

of purely real-valued and purely imaginary-valued wave functions are invariant subspaces of H and $\mathcal{P}^a$. If we define for $t \geq 0$ the spaces

$$\mathbb{L}^2 := \hat{\mathcal{L}}^2_{\mathbb{R}} := \hat{\mathcal{L}}^2 \cap \mathcal{L}^2_{\mathbb{R}}, \quad \mathbb{H}^t := \hat{\mathcal{H}}^t_{\mathbb{R}} := \hat{\mathcal{H}}^t \cap \mathcal{L}^2_{\mathbb{R}}, \quad \hat{\mathcal{L}}^2_{\mathbb{C}} := \hat{\mathcal{L}}^2 \cap \mathcal{L}^2_{\mathbb{C}}, \quad \mathcal{H}^t_{\mathbb{C}} := \hat{\mathcal{H}}^t \cap \mathcal{L}^2_{\mathbb{C}},$$

there holds

$$H|_{\hat{\mathcal{L}}^2} = \Big(H : \mathbb{H}^2 \to \mathbb{L}^2\Big) \oplus \Big(H : \hat{\mathcal{H}}^2_{\mathbb{C}} \to \hat{\mathcal{L}}^2_{\mathbb{C}}\Big). \tag{1.27}$$

(iii) Let S^z_N and $\mathcal{L}^2_k$, $k \in \{0, \ldots, N\}$, denote the z-spin operator and the spin spaces from Section 1.2(ii) respectively. There holds $HS^z_N\Psi = S^z_N H\Psi$ for all $\Psi \in \mathcal{H}^2$; H therefore maps $\mathcal{H}^2_k$ to $\mathcal{L}^2_k$.

(iv) Let

$$\mathbb{L}^2_k \ := \ \mathcal{L}^2_k \cap \mathbb{L}^2, \quad \mathbb{H}^t_k \ := \ \mathcal{L}^2_k \cap \mathbb{H}^t \tag{1.28}$$

for $t \geq 0$; the Hamiltonian $H : \mathbb{H}^2 \to \mathbb{L}^2$, i.e. restricted to the real-valued antisymmetric functions, can then be decomposed to a direct sum of Hamiltonians densely defined on the spin subspaces $\mathbb{L}^2_k$,

$$\big(H : \mathbb{H}^2 \to \mathbb{L}^2\big) \ = \ \bigoplus_{0 \leq k \leq N} \big(H : \mathbb{H}^2_k \to \mathbb{L}^2_k\big). \tag{1.29}$$

(v) For spin numbers $k, k' \in \{0, \ldots, N\}$ with $k + k' = N$ (i.e. $\zeta_k = -\zeta_{k'}$), the antisymmetrized spaces $\mathbb{L}^2_k, \mathbb{L}^2_{k'}$ are isomorphic by spin conjugation, i.e. by componentwise multiplication of the spin vector of Ψ with -1. The action of H on the real-valued antisymmetric space $\mathbb{H}^2$ is therefore already determined by its action on the spaces $\mathbb{H}^2_k$, $0 \leq k \leq \frac{N}{2}$.

Proof. All claims are straight-forwardly deduced from the structure of H, namely the facts that it is invariant under permutation of indices, only incorporates real quantities, and that it does not act on the spin variable, together with the simple observation that $\mathcal{P}^a$, S^z_N and c all map $\mathcal{H}^2 \to \mathcal{H}^2$ (for $\mathcal{P}^a$, also see Lemma 1.5).

□

1.3 The strong and the weak form of the electronic Schrödinger equation

(i) The strong eigenvalue equation. By 1.1(ii), the energy values an electronic system can attain are determined by the spectrum

$$\text{spec}(H) \;:=\; \{\, \lambda \in \mathbb{C} \mid H - \lambda I \;\text{ does not have a bounded inverse } \} \tag{1.30}$$

of the according Hamiltonian, itself governed by the form of the potential energy term V. While for some potentials like bounded potentials on bounded domains, the according Hamiltonian may have a compact resolvent, so that standard operator eigenvalue theory may be applied to show that its spectrum only consists of discrete eigenvalues [47], or while in other cases, a complete set of eigenfunctions may be calculated explicitly as for the quantum mechanical harmonic oscillator [214], those results do unfortunately not apply to the electronic Hamiltonian (1.20), and the electronic Hamiltonian H admits for a rather complicated spectrum.[14]

Let us denote by $\hat{\mathcal{L}}^2_b \subseteq (\mathcal{L}^2, \|.\|)$ the space spanned by the antisymmetric eigenvectors of H, i.e. by those antisymmetric functions $0 \neq \Psi \in \hat{\mathcal{H}}^2$ for which there is an $E \in \mathbb{C}$ such that the *eigenpair* (Ψ, E) fulfils the *time-independent electronic Schrödinger equation*

$$H\Psi = E\Psi. \tag{1.31}$$

By a result going back to Ruelle [183], sometimes termed the RAGE theorem [111, 202], $\hat{\mathcal{L}}^2_b$ is the space that contains all electronic states that remain localized for all times, therefore representing the electronic *bound states* of the given molecule; the corresponding expectation values are their corresponding energies. In particular, if H has any eigenvalues at all,

$$E_0 \;=\; \inf_{0 \neq \Psi \in \mathcal{H}^2 \cap \hat{\mathcal{L}}^2_b} \frac{\langle H\Psi, \Psi \rangle}{\langle \Psi, \Psi \rangle} \tag{1.32}$$

is an eigenvalue of H, representing the *electronic ground state energy* of the given molecule. The (approximate) computation of E_0 and a corresponding eigenvector is one of the central tasks of electronic structure calculation and of this work. Using (1.27), it is not hard to show that

$$\text{spec}(H : \hat{\mathcal{H}}^2 \to \hat{\mathcal{L}}^2) \;=\; \text{spec}(H : \mathbb{H}^2 \to \mathbb{L}^2).$$

From real-valued eigenfunctions, complex eigenfunctions are then constructed by taking the real-valued solutions (belonging to the same eigenvalue) as their real and imaginary part. In particular, an eigenvalue E is simple in $\hat{\mathcal{H}}^2$ iff it is simple in $\mathbb{H}^2$, and the lowest eigenvalue (1.32) of H coincides with the lowest eigenvalue belonging to a real-valued eigenfunction. Therefore, using Lemma 1.9, the computation of the ground state (1.32) amounts to the following first formulation of the central problem of this work.

[14] For results about the spectral properties of Hamiltonians with other potentials V, see [28, 179] and [197] and the extensive references therein.

Problem 1.10. *(Strong eigenvalue problem for H)*
Provided that the electronic Hamiltonian $H : \mathbb{H}^2 \to \mathbb{L}^2$ from (1.20) has a non-empty point spectrum, find (or approximate) an antisymmetric function $\hat{\Psi} \in \mathbb{H}^2$ such that it is an eigenfunction of H belonging to the lowest eigenvalue $E_0 \in \mathbb{R}$ that H attains on $\mathbb{H}^2$, that is, $\hat{\Psi}$ is a solution of the *time-independent Schrödinger equation*

$$H\hat{\Psi} \quad = \quad E_0\hat{\Psi}, \tag{1.33}$$

and E_0 fulfils (1.32).

□

(ii) The weak eigenvalue equation. For numerical treatment of partial differential equations, it is common practice to skip from the above strong formulation (1.33) to the weak formulation. This way, one circumvents the problems associated with the treatment of unbounded operators and obtains a natural way of discretising and analysing the corresponding equations, see e.g. [95] for an introduction.

Definition 1.11. *(Electronic Hamiltonian bilinear form and weak eigenpairs)*
For the Hamiltonian $H : \mathbb{H}^2 \to \mathbb{L}^2$ (restricted to the real-valued antisymmetric space $\mathbb{H}^2$), the associated symmetric bilinear form is given by

$$h : \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R}, \qquad h(\Psi, \Psi') := \langle H\Psi, \Psi' \rangle = \underbrace{\frac{1}{2}\langle \nabla\Psi, \nabla\Psi' \rangle}_{t(\Psi,\Psi')} \; + \; \underbrace{\langle V(x)\Psi, \Psi' \rangle}_{v(\Psi,\Psi')}. \tag{1.34}$$

The potential energy bilinear form v, and thus also h, can be extended to a continuous bilinear form on $\mathbb{H}^1 \times \mathbb{H}^1$ [211], which is also given explicitly by (1.34) and which we also denote by h. *Weak (electronic) eigenpairs* of h are pairs $(\Psi, E) \in \mathbb{H}^1 \times \mathbb{R}$ for which

$$h(\Psi, \Psi') = E\langle \Psi, \Psi' \rangle \quad \text{for all} \quad \Psi' \in \mathbb{H}^1. \tag{1.35}$$

It can be shown that a function $\Psi \in \mathbb{H}^1$ is a weak eigenfunction in the sense of (1.35) if and only if $\Psi \in \mathbb{H}^2$ and Ψ fulfills the strong Schrödinger equation (1.33), see [214]. Problem 1.10 is therefore equivalent to finding a "weak eigenpair"

$$(\hat{\Psi}, E_0) \in \mathbb{H}^1 \times \mathbb{R}$$

of h, where E_0 is the lowest eigenvalue of h, and we will go on by decomposing this problem further.

(iii) Decomposition of h. From Lemma 1.9, it follows that

$$\mathrm{spec}(H : \mathbb{H}^2 \to \mathbb{L}^2) \quad = \quad \bigcup_{0 \le k \le \frac{N}{2}} \mathrm{spec}(H : \mathbb{H}^2_k \to \mathbb{L}^2_k).$$

Therefore, we can (thanks to continuity arguments) accordingly decompose the weak eigenvalue problem into eigenvalue problems for fixed k, $0 \le k \le \frac{N}{2}$. Because there holds $H\Psi(X, \sigma) = 0$ for each spin vector σ for which $\sum_{s_i \in \sigma} s_i \neq -\frac{N}{2} + k$, we can thus reformulate Problem 1.10 in terms of an equivalent set of no more than $(N+1)/2$ problems, with which we will deal from now on:

Problem 1.12. *For fixed k with $0 \le k \le N/2$, find an eigenpair*

$$(\underline{\Psi}, E^*) \;=\; (\underline{\Psi}_k, E^*_k) \;\in\; \mathbb{H}^1_k \times \mathbb{R}$$

such that E^* is the lowest eigenvalue of the bilinear form h on $\mathbb{H}^1_k \times \mathbb{H}^1_k$, i.e.

$$h(\underline{\Psi}, \Psi') \quad = \quad E^* \langle \underline{\Psi}, \Psi' \rangle \qquad \textit{for all} \;\; \Psi' \in \mathbb{H}^1_k, \tag{1.36}$$

and E^* is the smallest value such that there is a $\underline{\Psi} \in \mathbb{H}^1_k$ for which (1.36) holds.

For each k, we can now compute an eigenpair $(\underline{\Psi}_k, E^*_k) \in \mathbb{H}^2_k \times \mathbb{R}$, where E^*_k is the lowest eigenvalue of h restricted in the left argument to $\mathbb{H}^2_k$. The overall ground state energy E_0 is then given by the lowest of those eigenvalues.

(iv) Existence of bound and ground states; lower bound for h.

If for fixed $0 \le k \le N/2$, the infimum

$$\inf\{\; h(\Psi, \Psi) \mid \Psi \in \mathbb{H}^1_k, \quad \|\Psi\| = 1\} \tag{1.37}$$

is an eigenvalue of h, Problem 1.12 is equivalent to computing the minimum and minimizer of (1.37), and thus to a classical minimization task. However, this does not necessarily need to be the case, and we may even encounter the situation that the lower eigenvalues of h are "hidden" in the essential spectrum [179] of h, making the computation of these eigenvalues a numerically tedious task. In the context of electronic structure calculation, we are offered a way out by the fact that the bottom $\inf \mathrm{spec}_{ess}$ of the essential spectrum can be associated with a formalization of the ionization threshold energy of the molecule (see e.g. [4, 167, 214]). Therefore, if we make assumption that for a configuration of fixed spin number $0 \le k \le N/2$, it is energetically more advantageous for the electrons to stay in the vicinity of the nuclei than to fade away at infinity (which seems physically

reasonable if we want to compute stable molecules), this assumption implies [214][15] the following statement, which we will assume from this point on.

Assumption 1.13. *For fixed z-spin value $\zeta_k \in spin(N)$, there holds that*

$$\mathcal{E}^* := \inf\{\ h(\Psi,\Psi) \mid \Psi \in \mathbb{H}^1_k, \quad \|\Psi\| = 1\ \} \quad < \quad \inf \mathrm{spec}_{ess}(h|_{\mathbb{H}^1_k \times \mathbb{H}^1_k}). \tag{1.38}$$

Under this condition, every value E contained in the spectrum of $h|_{\mathbb{H}^1_k \times \mathbb{H}^1_k}$ and smaller than the ionization energy $\inf \mathrm{spec}_{ess}(h|_{\mathbb{H}^1_k \times \mathbb{H}^1_k})$ is an eigenvalue of finite multiplicity, i.e. a bound state of the molecule; in particular, $E^* := \mathcal{E}^*$ is the lowest one, the *ground state energy* of the given nuclear configuration, and the corresponding eigenfunction Ψ is an *electronic ground state* of the configuration. Also, Assumption 1.13 vindicates the Rayleigh-Ritz principle, providing a solid basis for a variational analysis, see also [214]. Assumption 1.13 can be proven for some cases, e.g. for one-atomic molecules; for the case of $N = 2$, conditions on the decay of the potential may be given to enforce a finite discrete spectrum, for greater N, its validity may be related to the total charge for atoms. For a review of those and related results on the spectral properties of N-electron Hamiltonians and other cases and for the related HVZ-theorem, confer [197] or the quite exhaustive review [111] and the references therein.

The bilinear form $h : \mathbb{H}^1_k \times \mathbb{H}^1_k$, and therefore also $h : \mathbb{H}^1 \times \mathbb{H}^1$, can be shown to fulfil a Gårding inequality [209] on $\mathcal{H}^1$ [214]: There holds

$$c\ \|\Psi\|_1^2 - \mu\langle\Psi,\Psi\rangle \quad \leq \quad h(\Psi,\Psi) \quad \leq \quad C\ \|\Psi\|_1^2 \tag{1.39}$$

for all $\Psi \in \mathbb{H}^1$ and some $\mu \in \mathbb{R}$, $c, C > 0$. We will later use (1.39) to show that $h(\cdot,\cdot) - E_0\langle\cdot,\cdot\rangle$ is a bounded, $\mathbb{H}^1_k$-elliptic mapping on the orthogonal complement of the eigenspace belonging to $E_0 = \mathcal{E}_0$, an indispensable tool in the analysis of the algorithms of quantum chemistry, see Section 2.

(v) Properties of electronic eigenfunctions. We compile only some of the vast amount of known facts and references about properties of electronic eigenfunctions Ψ very briefly, and refer to [198] for a detailed review. Most results are formulated for the spatial components $\Psi_X := \Psi(\cdot,\sigma)$ with a fixed spin vector σ, which are in an obvious way related to the eigenfunctions $\Psi \in \mathbb{H}^1_k$ by the antisymmetry constraint. Ψ_X is for any $0 < \theta < 1$ (almost everywhere equal to) a θ-Hölder-continuous function on all of $\mathbb{R}^{3N}$ with locally bounded derivatives, so Ψ_X is (almost everywhere equal to) a locally Lipschitz

[15] In [214], the below condition is stated for certain antisymmetric subspaces $H(\sigma)$ of $L^2(\mathbb{R}^{3N})$ to which the weak eigenvalue problem may be decomposed for a nice analysis. It is not hard to show though that the condition from [214] and the one given below are equivalent.

continuous function [114]. If $E < \inf \mathrm{spec}_{ess}$, Ψ_X and its partial derivatives decay as $e^{-\sqrt{2(\Sigma - E)}|X|}$ for any value $E < \Sigma < \inf \mathrm{spec}_{ess}$ [4, 162], cf. also [214] for related results. Similar results also hold for so-called non-threshold eigenvalues lying in the essential spectrum of H, see [111]. From this, an according pointwise bound $|\Psi(X)| \leq Ce^{-\sqrt{2(\Sigma - E)}|X|}$ can be deduced by methods explained in [198]. Note that this refines the characterization of eigenfunctions as "bound states" of the system and also vindicates the computation of Ψ_X on $\mathbb{R}^{3N}$ by approximation on bounded domains. In particular, Ψ_X is bounded in $\mathbb{R}^{3N}$, see also [198].

Concerning regularity, we note at first that by the equivalence of weak and strong formulation, weak eigenfunctions are globally $\mathcal{H}^2$. Not much more can be expected globally, as already the simple example of the of the hydrogen atom, for which the ground state can be computed analytically, has a Sobolev regularity limited to $t < 5/2$. Nevertheless, standard results (see e.g. Theorem 8.10. in [85]) can be used to show that eigenfunctions Ψ are C^∞ at any $X \in \mathbb{R}^{3N}$ where $V(X)$ is C^∞. The complement of those points is the set of *coalescence points* or *cusps* of the wave function, where either $x_i = R_j$ for some $i \in N]\,, j \in M]$, or $x_i = x_j$ for some $i, j \in N]$, that is, where an electron and a nucleon meet or where at least two electrons are at the same place in space. For the behaviour of Ψ_X on the cusp set, the hydrogen atom provides an instructive example that shows that the derivative of Ψ_X does not need to be continuous at coalescence points. More general results were first formulated by Kato [114] for points where exactly two particles meet, and later extended to the general case in a series of papers [79, 106, 107, 108]. See [79] for a quite clear characterization of the cusps. In numerical computations, approximation of the electron-electron cusps ($x_i = x_j$) poses a major obstacle when employing the classical methods used in quantum chemistry. In this context, another interesting family of regularity results has recently been proven by Yserentant [212, 213, 214]: By using the antisymmetry condition, it can be shown that specific mixed first derivatives of Ψ and their derivatives exist, are square-integrable and decay exponentially. In particular, an N particle wave function enjoys increasing mixed regularity with increasing particle number N, even in the cusp points. Thus, the "curse of dimensionality", i.e. in this case the exponential dependence of the dimension of the discretisation of $\mathcal{L}^2_N$ on the number N of particles, can be broken at least theoretically by use of sparse grids techniques [38], cf. [92, 213, 216] for some results.

□

1.4 Bases for tensor spaces and the Slater basis

To discretise Problem 1.12 Galerkin-style, and also to formulate the eigenvalue problem in terms of Second Quantization, a basis of the antisymmetric, real valued tensor spaces $\mathbb{H}^1_k$ and $\mathbb{L}^2_k$ for fixed $0 \leq k \leq N/2$ is needed. The according constructions and notations are introduced in this section.

As a first step, we will construct a tensor basis $\mathcal{B}$ for the real spaces

$$\mathcal{L}^2_{\mathbb{R}} = L^2(\mathbb{R}^{3N} \times \Sigma^N, \mathbb{R}), \qquad \mathcal{H}^1_{\mathbb{R}} := \mathcal{H}^1 \cap \mathcal{L}^2_{\mathbb{R}}.$$

We will then restrict this basis to certain ordered tensor bases $\mathbb{B}'_k$ for the spin numbers $0 \leq k \leq N/2$, which are in a third step mapped to the so-called Slater bases $\mathbb{B}_k$ of the antisymmetric spaces $\mathbb{L}^2_k$ and $\mathbb{H}^1_k$ which we have to deal with when treating the weak eigenvalue problem (1.36).

Definition 1.14. *(Spatial, spin and tensor space bases)*
Let

$$B \;:=\; \{\varphi_p \in H^1(\mathbb{R}^3, \mathbb{R}) \mid p \in \mathbb{N}\} \tag{1.40}$$

be a basis of $H^1(\mathbb{R}^3, \mathbb{R})$ (consisting of so called *spatial orbitals* φ_p). From each spatial orbital φ_p, we construct two so-called *spin orbitals* $\chi_{\overline{p}}, \chi_{\underline{p}} \in B \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$ of respective spin $+\frac{1}{2}$ and spin $-\frac{1}{2}$ by setting

$$\begin{array}{llll} \chi_{\overline{p}}(x, \frac{1}{2}) & = \varphi_p(x), & \chi_{\overline{p}}(x, -\frac{1}{2}) & = 0, \\ \chi_{\underline{p}}(x, \frac{1}{2}) & = 0 & \chi_{\underline{p}}(x, -\frac{1}{2}) & = \varphi_p(x). \end{array}$$

and then letting

$$B^\Sigma = \{\chi_{\overline{p}},\ \chi_{\underline{p}} \mid p \in \mathbb{N}\}. \tag{1.41}$$

To index elements from B^Σ, we let

$$\mathcal{I}^+ := \{\overline{p} \mid p \in \mathbb{N}\}, \quad \mathcal{I}^- := \{\underline{p} \mid p \in \mathbb{N}\}, \quad \mathcal{I} := \mathcal{I}^+ \cup \mathcal{I}^-. \tag{1.42}$$

Finally, we define an according basis of $\mathcal{L}^2_{\mathbb{R}}$ by

$$\mathcal{B} := \{\otimes_{k=1}^N \chi_{P_k}(x_k, s_k) \;\mid\; P_1, \ldots, P_N \in \mathcal{I}\}. \tag{1.43}$$

Elements of $\mathcal{B}$, not specified further, will be labeled by a "tilde", $\tilde{\Psi} \in \mathcal{B}$.

□

Remarks 1.15.

(i) By the results on tensor product spaces in Section 1.1(i), $\mathcal{B}$ is a basis of $\mathcal{L}^2_{\mathbb{R}}$ and also a basis of $\mathcal{H}^1 \cap \mathcal{L}^2_{\mathbb{R}}$ because $\mathcal{H}^1_{\otimes} \supseteq \mathcal{C}^\infty_0$ (see Section 1.1(iii) for the definition) is dense in $\mathcal{H}^1$.

(ii) If B is L_2-orthonormal, B^Σ is an L_2-orthonormal basis, and $\mathcal{B}$ is an orthonormal basis with respect to the $\mathcal{L}^2_{\mathbb{R}}$-inner product.

(iii) If a function $\Psi \in \mathcal{B}$ contains exactly k functions χ_P with indices from $\mathcal{I}^+$, it is obviously an eigenfunction of the z-spin operator S^z_N, corresponding to an eigenvalue $\zeta_k = -\frac{N}{2} + k$. Therefore, $\mathcal{B}$ is an eigenbasis of S^z_N.

□

Before defining a basis for the antisymmetric spaces $\mathbb{L}^2_k$, we will need an intermediate step, in which we construct for $0 \leq k \leq N$ ordered basis sets $\mathbb{B}'_k$, spanning according ordered tensor spaces $\mathbb{L}^{2,\mathrm{ord}}$. In step (ii), we map them to basis sets $\mathbb{B}_k$ of the spaces $\mathbb{L}^2_k$ by use of the antisymmetry projector $\mathcal{P}^a$.

Definition 1.16. *(Ordered tensor product bases, Slater basis)*

(i) On the index set $\mathcal{I}$ from (1.42), we introduce an ordering by defining

$$\overline{p} < \overline{q}, \quad \underline{p} < \underline{q} \ \text{ iff } \ p < q; \qquad \overline{p} < \underline{q}$$

for all $p, q \in \mathbb{N}$. If no particular spin is designated for an index from the set $\mathcal{I}$, it will be denoted by a capital letter $P, Q, \ldots$.
For $k \in \{0, \ldots, N\}$, we define the ordered tensor bases

$$\mathbb{B}'_k \ := \ \left\{ \otimes_{i=1}^N \chi_{P_i} \mid P_1 < \ldots < P_k \in \mathcal{I}^+, \ P_{k+1} < \ldots < P_N \in \mathcal{I}^- \right\}, \quad (1.44)$$

and let

$$\mathbb{B}' \ := \ \dot\cup_{k=0,\ldots,N} \mathbb{B}'_k, \qquad \mathbb{L}^{2,\mathrm{ord}}_k := \overline{\mathrm{span}(\mathbb{B}_k)}^{\mathcal{L}^2}, \qquad \mathbb{L}^{2,\mathrm{ord}} := \overline{\mathrm{span}(\mathbb{B})}^{\mathcal{L}^2}.$$

Elements from $\mathbb{B}'$ (only turning up in this section) will be marked by a "prime", $\Psi'_\mu \in \mathbb{B}'$.

(ii) Using the antisymmetry projector $\mathcal{P}^a$ from (1.22), we define the mapping

$$\mathcal{Q} \ : \ \mathcal{L}^2_{\mathbb{R}} \to \mathbb{L}^2, \quad \mathcal{Q}\Psi = \sqrt{N!} \cdot \mathcal{P}^a \Psi. \quad (1.45)$$

and the *Slater bases*

$$\mathbb{B}_k := \{\Psi_\mu := \mathcal{Q}\Psi'_\mu \mid \Psi'_\mu \in \mathbb{B}'_k\}, \qquad \mathbb{B} := \dot\cup_{k=0,\ldots,N} \mathbb{B}_k. \quad (1.46)$$

The terminology introduced in Definition 1.16 is justified by the next lemma.

Lemma 1.17. *(Slater determinants, isometric property of $\mathcal{Q} : \mathbb{L}^{2,ord} \to \mathbb{L}^2$, Slater basis)*

(i) For each function $\tilde{\Psi}_\mu = \otimes_{i=1}^N \chi_{P_i}(x_i, s_i) \in \mathcal{B}$ for which two indices in μ coincide, there holds

$$\mathcal{Q}\tilde{\Psi}_\mu = 0. \tag{1.47}$$

If all indices in μ are mutually distinct, its image under $\mathcal{Q}$ is given by a so-called Slater determinant,

$$\Psi_\mu \;=\; \hat{\otimes}_{i=1}^N \chi_{P_i}(x_i, s_i) \;:=\; \mathcal{Q}\tilde{\Psi}_\mu \;=\; \frac{1}{\sqrt{N!}} \sum_{\pi \in S(N)} \otimes_{i=1}^N \chi_{P_i}(x_{\pi(i)}, s_{\pi(i)}). \tag{1.48}$$

In particular,

$$\mathcal{Q}\tilde{\Psi}_\mu \;=\; \mathcal{Q}\tilde{\Psi}_\nu \tag{1.49}$$

iff all the indices in μ and ν coincide (except for possibly different ordering).

(ii) The restriction of $\mathcal{Q}$ to $\mathbb{L}^{2,ord}$ is an $\mathcal{L}^2$-orthogonal isomorphism between $\mathbb{L}^{2,ord}$ and the antisymmetric space $\mathbb{L}^2$, i.e $\mathcal{Q}$ is continuous, one-to-one and onto, and for any

$$\Psi_1', \Psi_2' \;\in\; \mathbb{L}^{2,ord}, \quad \Psi_1 \;:=\; \mathcal{Q}\Psi_1', \quad \Psi_2 \;:=\; \mathcal{Q}\Psi_2',$$

there holds

$$\langle \Psi_1, \Psi_2 \rangle = \langle \Psi_1', \Psi_2' \rangle. \tag{1.50}$$

In particular, $\mathbb{B}'$ is an $\mathcal{L}^2$-orthonormal basis of $\mathbb{L}^{2,ord}$ iff $\mathbb{B}$ is an $\mathcal{L}^2$-orthonormal basis of $\mathbb{L}^2$.

(iii) $\mathcal{Q}$ maps $\mathbb{L}_k^{2,ord}$ onto $\mathbb{L}_k^2$ for every $k \in \{0, \ldots, N\}$, so that for $\mathbb{B}_k'$ orthonormal, $\mathbb{B}_k$ is an $\mathcal{L}^2$-orthonormal basis of $\mathbb{L}_k^2$.

Proof. Concerning (i), we only note that (1.47) and (1.48) follow directly from the definition of $\mathcal{P}^a$, while (1.49) is proven by writing out $\mathcal{Q}\tilde{\Psi}_\mu, \mathcal{Q}\tilde{\Psi}_\nu$ and using that $S(N)$ is invariant under left multiplication with the index permutation that takes μ to ν. To show (1.50) from (ii), we fix an orthonormal tensor basis $\mathbb{B}'$ of $\mathbb{L}^{2,\mathrm{ord}}$. For a basis functions $\Psi_\mu', \Psi_\nu' \in \mathbb{B}'$, the definition of $\mathcal{Q}$ shows that $\|\mathcal{Q}\Psi_\mu'\| = 1$ while by Lemma 1.5,

$$\langle \Psi_\mu, \Psi_\nu \rangle \;=\; \langle \mathcal{Q}\Psi_\mu', \mathcal{Q}\Psi_\nu' \rangle \;=\; \sqrt{N!}\, \langle \mathcal{P}^a \Psi_\mu', \Psi_\nu' \rangle.$$

It is not hard to see that from this, $\langle \Psi_\mu, \Psi_\nu \rangle = 0$ follows if $\nu \neq \mu$; thus, for any $\Psi' = \sum_{\mu \in \mathcal{M}} \alpha_\mu \Psi_\mu' \in \mathbb{L}^{2,\mathrm{ord}}$,

$$\|\mathcal{Q}\Psi'\|^2 \;=\; \langle \mathcal{Q}(\sum_{\mu \in \mathcal{M}} \alpha_\mu \Psi_\mu'),\; \mathcal{Q}(\sum_{\nu \in \mathcal{M}} \alpha_\nu \Psi_\mu') \rangle \;=\; \sum_{\mu \in \mathcal{M}} \sum_{\nu \in \mathcal{M}} \alpha_\mu \alpha_\nu \delta_{\mu,\nu} \;=\; \|\Psi'\|^2.$$

This shows that $\mathcal{Q}$ is $\mathbb{L}^2$-norm-preserving and in particular continuous and one-to-one, and that the functions from $\mathbb{B}$ are linearly independent. To show that $\mathcal{Q}$ is onto, which also proves that $\mathbb{B}$ is a basis of $\mathbb{L}^2$, we note that from Lemma 1.9(i), there follows $\mathbb{L}^2 = \mathcal{P}^a\mathcal{L}^2_{\mathbb{R}}$; thus, it suffices to show that for any $\Psi \in \mathcal{L}^2_{\mathbb{R}}$,

$$\mathcal{P}^a\Psi = \mathcal{Q}\Psi'$$

for some $\Psi' \in \mathbb{B}'$. To start with, we notice that for any function $\tilde{\Psi}_\nu$ from the basis $\mathcal{B}$ of $\mathcal{L}^2_{\mathbb{R}}$, there either holds $\mathcal{P}^a\Psi_\nu = 0$, or there is a $\Psi'_{\pi_\nu} \in \mathbb{B}'$ for which

$$\mathcal{Q}\Psi'_{\pi_\nu} = \sqrt{N!}\,\mathcal{P}^a\tilde{\Psi}_\nu.$$

Indeed, let $\tilde{\Psi}_\nu := \otimes_{n=1}^N \chi_{P_n} \in \mathcal{B}$; then, if $\mathcal{Q}\Psi_\nu \neq 0$, all indices of ν are distinct, so there is a permutation π_ν of the basis functions χ_{P_n} such that

$$\Psi'_{\pi_\nu} \;:=\; \otimes_{k=1}^N \chi_{P_{\pi_\nu(n)}} \in\; \mathbb{B}$$

(namely the one sorting the indices according to the ordering on I). Because the symmetric group is invariant under right multiplication with the permutation π_ν, there follows

$$\mathrm{sgn}(\pi_\nu)\;\mathcal{P}^a\tilde{\Psi}_\nu \;=\; \mathcal{P}^a\Psi'_{\pi_\nu} \;=\; \frac{1}{\sqrt{N!}}\;\mathcal{Q}\Psi'_{\pi_\nu}.$$

Let us now denote by $\mathcal{N}^*$ the set multi-indices $\mu \in \mathcal{I}^N$ for which all indices are distinct. Because $\mathbb{L}^2 = \mathcal{P}^a\mathcal{L}^2_{\mathbb{R}}$, there is for any $\Psi \in \mathbb{L}^2_k$ a sequence $(\alpha_\nu)_{\nu\in\mathbb{N}}$ and a corresponding sequence of elementary tensors $\Psi_\nu \in \mathcal{B}$ such that

$$\Psi \;=\; \mathcal{P}^a\big(\sum_{\nu\in\mathcal{N}} \alpha_\nu\tilde{\Psi}_\nu\big) \;=\; \frac{1}{\sqrt{N!}}\sum_{\nu\in\mathcal{N}^*} \mathrm{sgn}(\pi_\nu)\alpha_\nu\mathcal{Q}\Psi'_{\pi_\nu} \;=\; \mathcal{Q}\big(\frac{1}{\sqrt{N!}}\sum_{\nu\in\mathcal{N}^*} \mathrm{sgn}(\pi_\nu)\alpha_\nu\Psi'_{\pi_\nu}\big),$$

and the rightmost expression is an element of $\mathcal{Q}\mathbb{L}^{2,\mathrm{ord}}$, showing that $\mathcal{Q} : \mathbb{L}^{2,\mathrm{ord}} \to \mathbb{L}^2$ is onto $\mathbb{L}^2$. Because also $\mathbb{L}^2_k = \mathcal{P}^a\mathcal{L}^2_k$ by Lemma 1.9(iii), an analogous argument shows that $\mathcal{Q}$ maps $\mathbb{L}^{2,\mathrm{ord}}_k$ to $\mathbb{L}^2_k$ and thus proves (iii).

□

Remark/Definition 1.18. (Index sets for the bases $\mathbb{B}, \mathbb{B}_k$)
Each Slater determinant $\Psi_\mu = \mathcal{Q}\Psi'_\mu \in \mathbb{B}$ is by the last lemma also uniquely labeled by a multi-index $\mu = (P_1, \ldots, P_N)$ from the set

$$\mathcal{M} \;:=\; \{\; \mu = (P_1, \ldots, P_N) \in \mathcal{I}^N \;\;|\;\; P_1 < \ldots < P_N \;\}. \tag{1.51}$$

If $\Psi_\nu \in \mathbb{B}_k$ for fixed spin index k, there holds

$$\mu \in \mathcal{M}_k \;:=\; \{\mu = (P_1, \ldots, P_N) \in \mathcal{M} \;\;|\;\; P_1, \ldots, P_k \in \mathcal{I}^+, P_{k+1}, \ldots, P_N \in \mathcal{I}^-\}, \tag{1.52}$$

so that

$$\mathbb{B} = \{\Psi_\mu \mid \mu \in \mathcal{M}\}, \qquad \mathbb{B}_k = \{\Psi_\mu \mid \mu \in \mathcal{M}_k\}. \tag{1.53}$$

□

1.5 The electronic Schrödinger equation in Second Quantization

In various methods used in quantum chemistry, including the Coupled Cluster method to be treated in Section 3, the use of the formalism of Second Quantization [27] greatly simplifies the derivation of implementable equations. In Second Quantization, operators defined on the antisymmetric tensor space $\mathbb{L}^2$ are written in terms of annihilation and creation operators belonging to a fixed one particle spin basis of $L^2(\mathbb{R}^3 \times \Sigma)$, inducing a tensor basis of $\mathbb{L}^2$ as constructed in the last section. Operators are then completely determined by a corresponding set of coefficients, see [206] for results on the related concept of "matrix operators". In this section, we will introduce annihilation and creation operators in part (i), leading in part (ii) to a mathematically rigorous definition of the (weak) Second Quantization Hamiltonian that will be used later.

(i) Annihilation and creation operators. We will in this part (i) have to utilize the antisymmetric, real valued space $\mathbb{L}^2 = \mathbb{L}^2_N$ for a varying number N of electrons. Therefore, the spaces, operators etc. under consideration will be equipped with an index N indicating the number of particles where needed. Because notations used are intuitive and only needed in this part, we will not introduce them at all length. From part (ii) on, the particle number N will be fixed again; consequently, the indices will be omitted again. Let us introduce the (fermion) Fock space [77]

$$\mathbb{F} := \bigoplus_{N=0}^{\infty} \mathbb{L}^2_N,$$

where the symbol $\bigoplus$ denotes the direct orthogonal sum of the antisymmetric N-fold tensor product Hilbert spaces $\mathbb{L}^2_N$. In $\mathbb{F}$, we may embed any N-electron state vector $\Psi_N \in \mathbb{L}^2_N$ by writing it as $(\delta_{k,N}\Psi_N)_{k\in\mathbb{N}} = (0, 0, \ldots, 0, \Psi_N, 0, \ldots)$. Note that the case $N = 0$ is also included in the above definition of the space $\mathbb{F}$. For this case, $\mathbb{L}^2_0$ is (by definition of the tensor product) the underlying field of the complex numbers. This is a one-dimensional vector space, thus containing up to a phase factor only one state vector called the *vacuum state* $|\rangle$. This state is in some sense the starting point for the formalism of second quantization, as any state vector may be created from it by the use of the creation operators introduced below.
Motivated by our application in Section 3, the following definition of those operators also allows for non-orthogonal basis sets and functions f not contained in the basis B^Σ; the naming of the operators introduced will be motivated in the remarks given afterwards.

Definition 1.19. *(Creation and annihilation operators)*

(i) For $1 \leq N \in \mathbb{N}$, $f \in L^2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$ and $\Psi_\mu \in \mathbb{B}_N$, we at first define

$$a^\dagger_{f,N}\Psi_\mu \quad := \quad \mathcal{Q}_{N+1}\big(f \otimes \Psi_\mu\big), \tag{1.54}$$

where

$$\mathcal{Q}_{N+1} : \mathcal{L}^2_{\mathbb{R},N+1} \to \mathbb{L}^2_{N+1}$$

is the mapping from Definition 1.16.
By linear continuation of the above definition to linear combinations, and by closing [206] the operator in $\mathbb{L}^2_N$, we extend[16] each $a^\dagger_{f,N}$ to a linear map

$$a^\dagger_{f,N} : \mathbb{L}^2_N \to \mathbb{L}^2_{N+1}.$$

For $N = 0$, we let $a^\dagger_{f,0}|\rangle = f \in \mathbb{L}^2_1$. The *creation operator* or *creator* of f is now defined by

$$a^\dagger_f : \mathbb{F} \to \mathbb{F}, \quad a^\dagger_f := \bigoplus_{N=0}^{\infty} a^\dagger_{f,N}. \tag{1.55}$$

In particular, if $f = \chi_P$ from the fixed basis set $\mathbb{B}$, we will denote $a^\dagger_P := a^\dagger_{\chi_P}$ for convenience.

(ii) We define the *annihilation operator* or *annihilator* $a_f : \mathbb{F} \to \mathbb{F}$ of f as the adjoint of the creation operator $a^\dagger_f : \mathbb{F} \to \mathbb{F}$ of f. The annihilator of a basis function $\chi_P \in \mathbb{B}$ is denoted by a_P.

□

We remark that for any normed finite linear combination $\Psi = \sum_{n=1}^{M} \alpha_\mu \Psi_\mu$ of basis functions, it is easy to show $\|a^\dagger_{f,N}\Psi\|_{\mathcal{L}^2} \leq \|f\|_{L^2}$, so (as was already asserted above,) the closure [206] of $a^\dagger_{f,N}$ is an operator $\mathbb{L}^2_N \to \mathbb{L}^2_{N-1}$.
Additionally, because the creation operator $a^\dagger_f$ is closed, the adjoint of the adjoint of $a^\dagger_f$ is $a^\dagger_f$, so that the adjoint of the annihilator a_f is indeed $a^\dagger_f$, as indicated by the notation.

[16] See the remarks after this definition.

Later on, we will need the properties of the annihilation and creation operators compiled in the following lemma. The proofs can - given in the so-called "ket notation"[17] - be found in [103, 201] or are generalized from them straightforwardly, so they are omitted here.

Lemma 1.20. *(Properties of the creation and annihilation operators)*

(i) For $f \in \operatorname{span}\{\chi_{P_1}, \dots \chi_{P_N}\}$, we have

$$a_f^\dagger\big(\hat{\otimes}_{n=1}^N \chi_{P_n}\big) = 0.$$

(ii) The action of a_f on an N-electron elementary tensor $\Psi = \otimes_{i=1}^N \chi_{P_i}$ is given by

$$\tilde{a}_f \Psi \quad := \quad \sum_{n=1}^{N} (-1)^{n-1} \langle f, \chi_{P_n} \rangle \; \mathcal{Q}\Big(\big(\otimes_{i=1}^{n-1} \chi_{P_i} \big) \otimes \big(\otimes_{i=n+1}^{N} \chi_{P_i} \big) \Big). \tag{1.56}$$

(iii) In particular, there holds for $\Psi_\mu = \otimes_{i=1}^N \chi_{P_i} \in \mathbb{B}$ and $P_i \in \{P_1, \dots, P_N\}$ that

$$a_{P_i,N}\big(\hat{\otimes}_{n=1}^N \chi_{P_n}\big) \quad = \quad (-1)^{i-1} \mathcal{Q}\Big(\big(\otimes_{n=1}^{i-1} \chi_{P_n} \big) \otimes \big(\otimes_{n=i+1}^{N} \chi_{P_n} \big) \Big) \quad \in \quad \mathbb{L}^2_{N-1},$$

so that a_{P_i} "annihilates" the basis function χ_{P_i} and adds a corresponding sign.

(iv) For $J \notin \{P_1 \dots P_N\}$,

$$a_J\big(\hat{\otimes}_{n=1}^N \chi_{P_n}\big) = 0,$$

where 0 is the zero vector $0 \in \mathbb{F}$ (not to be confused with the vacuum state).

(v) Using the anticommutator $[A, B]_+ = AB + BA$, there hold the anticommutator relations

$$[a_f, a_g]_+ = 0, \qquad [a_f^\dagger, a_g^\dagger]_+ = 0, \tag{1.57}$$

and if $f, g \in L^2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$ are orthogonal,

$$[a_f, a_g^\dagger]_+ = [a_f^\dagger, a_g]_+ = 0. \tag{1.58}$$

If B is an orthogonal one-electron basis,

$$[a_P, a_Q^\dagger]_+ = [a_P^\dagger, a_Q]_+ = \delta_{P,Q} \tag{1.59}$$

for all $P, Q \in \mathcal{I}$, where $\delta_{P,Q} = 1$ only if $P = Q$ and $\delta_{P,Q} = 0$ otherwise. Furthermore, all creation and annihilation operators are nilpotent,

$$a_f a_f = a_f^\dagger a_f^\dagger = 0. \tag{1.60}$$

The importance of creation and annihilation operators is rooted in the fact that any linear operator on $\mathbb{F}$ may be written as a sum of polynomials in creation and annihilation operators $a_I^\dagger, a_I$ [44]. In particular, this of course includes the Hamiltonian, and its second quantization form will be introduced in the next section.

[17] In quantum chemistry, Slater determinants are usually denoted in the ket notation $|P_1, \dots, P_N\rangle$, related to the above by $|P_1, \dots, P_N\rangle := \mathcal{Q}\Psi_\mu$ for any (not necessarily sorted) index $\mu = (P_1, \dots, P_N)$.

(ii) The weak Hamiltonian of Second Quantization. For numerical treatment of the Schrödinger equation, one usually fixes a basis $\mathbb{B}_k$ of $\mathbb{H}^1_k$ as constructed in Definition 1.16. For this basis (or rather for a finite selection from $\mathbb{B}_k$ in practice), the matrix elements $h(\Psi_\mu, \Psi_\mu)$ of the bilinear form h then have to be evaluated. By definition of h, this task involves for each pair Ψ_μ, Ψ_ν of Slater determinants with coinciding spin a high-dimensional integration over $\mathbb{R}^{3N}$, which would in view of the size of the tensor basis and additionally the dimension of the integration domain quickly become an infeasible task even for very small N. It is therefore an essential fact that in an orthonormal basis set, this task reduces due to the structure of the Hamiltonian to the computation of $\mathcal{O}(|D|^4)$ integrals, where $|D|$ is the size of the used discretised one particle basis set $\{\chi_p | p \in D \subseteq I\}$. Additionally, those integrals are now involving at most 2 spatial variables x_i, x_j, i.e. they are integrals over $\mathbb{R}^6$. We now introduce notations for those integrals, and afterwards derive the weak Hamiltonian of Second Quantization.

Definition 1.21. *(Antisymmetric integrals of quantum chemistry)*
For $\chi_P, \chi_Q, \chi_R, \chi_S \in B^\Sigma$, we introduce the *single electron interaction integrals*

$$h_{P,Q} := \frac{1}{2}\langle \nabla\chi_P, \nabla\chi_Q\rangle + \sum_{\nu=1}^{K}\langle \chi_P, \frac{Z_\nu}{|x_i - R_\nu|}\chi_Q\rangle \tag{1.61}$$

and the *electron pair interaction integrals*[18,19]

$$\langle PQ|RS\rangle := \sum_{s,s'\in\{\pm\frac{1}{2}\}} \int_{\mathbb{R}^6} \chi_P(x,s)\chi_Q(y,s')\frac{1}{|x-y|}\chi_R(x,s)\chi_S(y,s')\, dxdy \tag{1.62}$$

as well as the *antisymmetrized integrals*

$$\langle PQ\|RS\rangle := \langle PQ|RS\rangle - \langle PQ|SR\rangle. \tag{1.63}$$

□

[18]The notation for electron pair interaction integrals introduced here the is the standard physicist's notation for the Coulomb integrals, which may be read as abbreviation for the inner product in (1.62). Note though that concurrently to this, the so-called Mullikan notation $(PR\|QS)$ is preferred by most chemists, related to the above by $(PR\|QS) = \langle PQ\|RS\rangle$. To avoid confusion, we will stick to the physicist's notation in this work.

[19]Note that (except for the case of closed shell calculations, i.e. $k = N/2$) the integrals depend not only on the indices p, q, r, s for the spin free basis functions, but on the spin orbital indices P, Q, R, S, i.e. e.g. $\langle \underline{p}Q\|RS\rangle \neq \langle \overline{p}Q\|RS\rangle$ in general.

With these definitions at hand, we can now introduce the Second Quantization Hamiltonian.

Lemma 1.22. *(Second Quantization Hamiltonian)*
By standard functional analysis [206], the bilinear form $h : \mathbb{H}^1_k \times \mathbb{H}^1_k$ defines a corresponding bounded linear operator $\hat{H} : \ \mathbb{H}^1_k \to \mathbb{H}^{-1}_k$, which maps $\Psi \in \mathbb{H}^1_k$ to a functional

$$\hat{H}\Psi : \ \mathbb{H}^1_k \to \mathbb{R}, \quad \Psi' \mapsto h(\Psi, \Psi'). \tag{1.64}$$

If B from (1.40) is an L_2-orthonormal basis set, this operator is in terms of annihilation and creation operators given by

$$\hat{H} \ = \ \sum_{P,Q \in \mathcal{I}} h_{P,Q} a^\dagger_P a_Q \ + \ \frac{1}{2} \sum_{P,Q,R,S \in \mathcal{I}} \langle PQ \| RS \rangle a^\dagger_P a^\dagger_Q a_S a_R. \tag{1.65}$$

Proof. Because of the linearity and continuity of h on $\mathbb{H}^1_k$, it suffices to show the claim for all Slater basis function $\Psi_\mu = \otimes_{n=1}^N \chi_{Q_n}, \Psi_\nu = \otimes_{n=1}^N \chi_{P_n} \in \mathbb{H}^1_k$. The conjecture thus is a consequence of the following equalities, see below for some comments.

$$\begin{aligned}
h(\Psi_\mu, \Psi_\nu) &= \sum_{i=1}^N h_{P_i,Q_i} \big(\prod_{\ell \neq i} \langle \chi_{Q_\ell}, \chi_{P_\ell} \rangle \big) + \sum_{i,j=1}^N \langle P_i P_j \| Q_i Q_j \rangle \big(\prod_{\ell \neq i,j} \langle \chi_{Q_\ell}, \chi_{P_\ell} \rangle \big) \\
&= \sum_{i=1}^N h_{P_i,Q_i} \langle a_{Q_i} \Psi_\mu, a_{P_i} \Psi_\nu \rangle + \sum_{i,j=1}^N \langle P_i P_j \| Q_i Q_j \rangle \langle a_{Q_j} a_{Q_i} \Psi_\mu, a_{P_j} a_{P_i} \Psi_\nu \rangle \\
&= \sum_{P,Q \in \mathcal{I}} h_{P,Q} \langle a_Q \Psi_\mu, a_P \Psi_\nu \rangle + \frac{1}{2} \sum_{P,Q,R,S \in \mathcal{I}} \langle PQ \| RS \rangle \langle a_S a_R \Psi_\mu, a_Q a_P \Psi_\nu \rangle \\
&= \sum_{P,Q \in \mathcal{I}} h_{P,Q} \langle a^\dagger_P a_Q \Psi_\mu, \Psi_\nu \rangle + \frac{1}{2} \sum_{P,Q,R,S \in \mathcal{I}} \langle PQ \| RS \rangle \langle a^\dagger_P a^\dagger_Q a_S a_R \Psi_\mu, \Psi_\nu \rangle \\
&= \langle \hat{H} \Psi_\mu, \Psi_\nu \rangle.
\end{aligned}$$

In the preceding, the representation of h in the first line follows from evaluation of $h(\Psi_\mu, \Psi_\nu)$ for the antisymmetric Ψ, Ψ'. As this is rather straightforward, we do not prove it here for sake of brevity; see [201] for the related Slater-Condon rules. The transition from the first to the second third line is due to (iii) of Lemma 1.20, while the third follows from (iv) of Lemma 1.20 and the fourth from the adjoint relation between a_I and $a^\dagger_I$. Additionally, symmetry of the coefficients and orthogonality of the basis functions were used.

□

1.6 Ellipticity results for the Hamiltonian and for Hamiltonian-like operators

For our analysis of the methods and algorithms of Quantum Chemistry in the next three sections, we will often use that the different operators turning up in the respective context are elliptic, or can be shifted to be elliptic on certain subspaces of the space under consideration. The present Section 1.6 compiles some results needed later.

We start with a simple lemma. The conditions of the essential Corollary 1.24 deduced from it are for example fulfilled by the Second Quantization Hamiltonian $\hat{H}$, see (1.39), and by some Fock- and Kohn-Sham type operators defined later, see Remark 2.8 and also Remark 3.14. For further conditions under which (1.66) holds, see [209].

Lemma 1.23. *Let $V \hookrightarrow X \hookrightarrow V'$ be a Gelfand triple, and let $A : V \to V'$ be a symmetric operator which is bounded from below by a Gårding estimate*

$$\langle Av, v\rangle \quad \geq \quad c_1\|v\|_V^2 \; - \; c_2\|v\|_X^2 \tag{1.66}$$

with constants $c_1, c_2 > 0$. If additionally, A is X-elliptic, i.e.

$$\langle Av, v\rangle \quad \geq \quad c_3\|v\|_X^2 \quad \textit{for all} \quad v \in V \tag{1.67}$$

for some $c_3 > 0$, then A is also V-elliptic,

$$\langle Av, v\rangle \quad \geq \quad c_4\|v\|_V^2 \quad \textit{for all} \quad v \in V \tag{1.68}$$

for some $c_4 > 0$.

Proof. For $q := c_3/(c_2 + c_3) < 1$, we use that $\langle Av, v\rangle$ can be expressed as

$$q\langle Av, v\rangle \; + \; (1-q)\langle Av, v\rangle \quad \geq \quad qc_1\|v\|_V^2 \; + \; (c_3 - q(c_2 + c_3))\|v\|_X^2 \quad = \quad qc_1\|v\|_V^2.$$

□

Corollary 1.24. *Let $V \hookrightarrow X \hookrightarrow V'$ be a Gelfand triple, and $A : V \to V'$ a symmetric operator with its lowest eigenvalue λ of finite multiplicity and bounded away from the rest of the spectrum of A,*

$$\lambda \quad < \quad \Lambda^* \; := \; \inf \; (\mathrm{spec}(A)\backslash\{\lambda\}).$$

Then, if A fulfils (1.66) of the previous lemma, $A - \lambda I$ is V-elliptic on the complement $V_\lambda^\perp$ of the eigenspace belonging to λ, i.e. there holds for some $c > 0$ that

$$\langle (A - \lambda I)v, v\rangle \quad \geq \quad c \; \|v\|_V^2 \quad \textit{for all} \quad v \in V_\lambda^\perp. \tag{1.69}$$

Proof. We apply Lemma (1.23) to the (symmetric) restriction of $A - \lambda I$ to the space $V_\lambda^\perp$, where thanks to the Courant-Fischer theorem [179],

$$\langle (A - \lambda I)v, v\rangle \quad \geq \quad (\Lambda^* - \lambda)\ \|v\|_X^2.$$

□

Next, we show that a norm $\|.\|_F$ equivalent to the norm on $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$ can be "lifted" to obtain a norm equivalent to that on $\mathcal{H}^1_{\mathbb{R}}$.

Lemma 1.25. *(Induced one-particle norm)*

Let

$$F: \ H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R}) \ \to \ H^{-1}(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$$

be a symmetric, bounded and elliptic linear mapping, i.e. with $\|\cdot\|_1$ denoting the norm on $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$,

$$\gamma\ \|\varphi\|_1^2 \quad \leq \quad \langle F\varphi, \varphi\rangle \quad \leq \quad \Gamma\ \|\varphi\|_1^2 \tag{1.70}$$

for some constants $\gamma, \Gamma > 0$ and all $\varphi \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$.

(a) The bilinear induced on $\mathcal{H}^1$ by F_N (cf. Def. 1.1), given in terms of the basis functions by

$$\langle \Psi_\mu, \Psi_\nu\rangle_F \quad := \quad \sum_{i=1}^{N} \langle \chi_{P_i}, \chi_{Q_i}\rangle_F \big(\prod_{j\neq i} \langle \chi_{P_j}, \chi_{Q_j}\rangle\big), \tag{1.71}$$

defines an inner product on $\mathcal{H}^1_{\mathbb{R}}$ for which the norm equivalence

$$\|\Psi\|_1 \quad \sim \quad \|\Psi\|_F \tag{1.72}$$

holds for all $\Psi \in \mathcal{H}^1_{\mathbb{R}}$.

(b) If $\epsilon > 0$ is a lower bound for the spectrum of F, then

$$\langle \Psi, \Psi\rangle_F \quad \geq \quad N\epsilon \|\Psi\|^2 \tag{1.73}$$

for all $\Psi \in \mathcal{H}^1_{\mathbb{R}}$.

□

Proof. We choose an orthonormal tensor basis $\mathcal{B}$ (constructed along the lines of Definition 1.14). We now show (1.72) for any finite linear combination $\Psi = \sum_{\mu\in\mathcal{N}} t_\mu \Psi_\mu$ of basis functions (i.e. only finitely many coefficients t_μ are nonzero); the assertions then follow by standard arguments. For each pair of basis functions $\Psi_\mu = \chi_1^\mu \otimes \ldots \otimes \chi_N^\mu$, $\Psi_\nu = \chi_1^\nu \otimes \ldots \otimes \chi_N^\nu$ with $\chi_i^\mu, \chi_i^\nu \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$, $i \in N]$, their $\mathcal{H}^1$-inner product (1.16) fulfils

$$\langle \Psi_\mu, \Psi'_\nu \rangle_1 \quad \sim \quad \sum_{m=1}^{N} \Big(\langle (1+|\omega|^2)(\mathcal{F}\chi_m^\mu)(\omega), (\mathcal{F}\chi_m^\nu)(\omega) \rangle \big(\prod_{l\neq m} \langle \chi_l^\mu, \chi_l^\nu \rangle \big) \Big)$$

by definition of the Fourier transform and a simple estimate, so that

$$\langle \Psi, \Psi \rangle_1 \quad \sim \quad \sum_{m=1}^{N} \Big(\sum_{\mu,\nu\in\mathcal{I}^N} \langle (1+|\omega|^2)\mathcal{F}\chi_m^\mu, \mathcal{F}\chi_m^\nu \rangle \big(\prod_{l\neq m} \langle \chi_l^\mu, \chi_l^\nu \rangle \big) \Big). \tag{1.74}$$

We will now estimate the term corresponding to a fixed $m \in N]$, and suppose $m = 1$ without of loss of generality. Let us denote by $\mathcal{N}^- = \mathcal{I}^{N-1}$ the multi-indices of length $N-1$. Define for any $\hat{\mu} \in \mathcal{N}^-$ a function

$$\chi_{\hat{\mu}} := \sum_{\substack{\mu=(I_1,\ldots,I_N)\in\mathcal{I}^N \\ (I_2,\ldots,I_N)=\hat{\mu}}} t_\mu \chi_{I_1}.$$

Then, because in (1.74), the rightmost product only is nonzero if the indices $I_2, \ldots, I_N$ of μ and ν coincide, we can contract the functions with coinciding second to last indices in the first component of the tensor product,

$$\begin{aligned} \sum_{\mu,\nu\in\mathcal{N}^-} \langle (1+|\omega|^2)\mathcal{F}\chi_1^\mu, \mathcal{F}\chi_1^\nu \rangle \Big(\prod_{2\leq l\leq N} \langle \chi_l^\mu, \chi_l^\nu \rangle \Big) \quad &= \quad \sum_{\hat{\mu}\in\mathcal{N}^-} \langle (1+|\omega|^2)\mathcal{F}\chi_{\hat{\mu}}, \mathcal{F}\chi_{\hat{\mu}} \rangle \\ = \quad \sum_{\hat{\mu}\in\mathcal{N}^-} \|\chi_{\hat{\mu}}\|^2_{H^1(\mathbb{R}^3)} \quad \sim \quad \sum_{\hat{\mu}\in\mathcal{N}^-} \langle F\chi_{\hat{\mu}}, \chi_{\hat{\mu}} \rangle \quad &= \quad \sum_{\mu,\nu\in\mathcal{N}^-} \langle F\chi_1^\mu, \chi_1^\nu \rangle \Big(\prod_{1\leq l\leq N} \langle \chi_l^\mu, \chi_l^\nu \rangle \Big), \end{aligned}$$

giving the first component of the F-inner product. analogous arguments hold for $m = 2, \ldots, N$, so the claim for (a) follows. Using $\langle F\chi_{\hat{\mu}}, \chi_{\hat{\mu}} \rangle \geq \epsilon \langle \chi_{\hat{\mu}}, \chi_{\hat{\mu}} \rangle$, the proof for (b) is very similar, so it is omitted.

□

For the so-called Fock operator (to which the above abbreviation F was an allusion, see Section 2), Lemma 1.23 can be used to show that (1.70) is fulfilled, and the previous Lemma 1.25 will be helpful in this context to prove some theoretical results using the shifted Fock operator as preconditioner. Unfortunately, the shift parameter can only be estimated. The following lemma shows that if a spin-wise HOMO-LUMO gap condition is fulfilled, shifting the lifted operator F_N by a sum of lowest eigenvalues taken in the spin components gives an elliptic operator $F_N - \Lambda_0 I$, which has the advantage of being computable for preconditioning in practice.[20]

[20]The assumptions of Lemma 1.26 reflect - except for the HOMO-LUMO condition (1.77) - the general Assumption 3.1 of Section 3. Because the Lemma is to the author's mind a little off-topic in Section 3, and proof is surprisingly technical and lengthy due to spin and antisymmetry issues, we decided to outsource it to the present section comprising technical results.

Lemma 1.26. *Let $F : H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R}) \to H^{-1}(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}, \mathbb{R})$ be a symmetric operator fulfilling the Gårding estimate (1.66). For fixed $k \in \{0, \ldots, N\}$, let*

$$\begin{aligned} M^+ &= \{P_1, \ldots, P_k\} \subseteq \mathcal{I}^+, & M^- &= \{P_{k+1}, \ldots, P_N\} \subseteq \mathcal{I}^+, \\ \overline{M}^+ &= \mathcal{I}^+ \backslash M^+, & \overline{M}^- &= \mathcal{I}^+ \backslash M^-, \end{aligned}$$

and let the spaces

$$\begin{aligned} V^+ &= \overline{\operatorname{span}\{\chi_P \mid P \in M^+\}}^{\|\cdot\|_1}, & V^- &= \overline{\operatorname{span}\{\chi_P \mid P \in M^-\}}^{\|\cdot\|_1}, \\ \overline{V}^+ &= \overline{\operatorname{span}\{\chi_P \mid P \in \overline{M}^+\}}^{\|\cdot\|_1}, & \overline{V}^- &= \overline{\operatorname{span}\{\chi_P \mid P \in \overline{M}^-\}}^{\|\cdot\|_1} \end{aligned}$$

be chosen such that

$$V^+ \perp_{L_2} \overline{V}^+, \quad V^+ \perp_F \overline{V}^+, \quad V^- \perp_{L_2}, \overline{V}^-, \quad V^- \perp_F \overline{V}^- \tag{1.75}$$

and such that F fulfils the spin-wise HOMO-LUMO gap conditions

$$\overline{\lambda} := \max_{\chi \in V^+} \langle F\chi, \chi\rangle < \inf_{\chi \in \overline{V}^+} \langle F\chi, \chi\rangle =: \overline{\Lambda}^*, \tag{1.76}$$

$$\underline{\lambda} := \max_{\chi \in V^-} \langle F\chi, \chi\rangle < \inf_{\chi \in \overline{V}^-} \langle F\chi, \chi\rangle =: \underline{\Lambda}^*. \tag{1.77}$$

Then, for $V = V^+ \oplus V^-$, $\Lambda_0 := tr(F|_V)$ is a (weak) simple eigenvalue of the lifted operator

$$F_N : \quad \mathbb{H}^1_k \to \mathbb{H}^{-1}_k,$$

and $F_N - \Lambda_0 I$ is a $\mathbb{H}^1_k$-elliptic operator on the orthogonal complement $U_0^\perp$ of span$\{\hat{\otimes}_{i=1}^N \chi_{P_i}\}$, i.e. there holds

$$\langle \Psi, (F - \Lambda_0)\Psi\rangle \geq \gamma \, \|\Psi\|^2_{\mathbb{H}^1} \tag{1.78}$$

for all $\Psi \in U_0^\perp$ and some $\gamma > 0$.

Proof. We define

$$\Psi_0 = \hat{\otimes}_{i=1}^N \chi_{P_i}, \qquad \mathbb{B}^*_k := \{\Psi_\mu \mid \Psi_\mu \in \mathbb{B}_k, \Psi_\mu \neq \Psi_0\};$$

thus, we have $U_0^\perp = \operatorname{span} \mathbb{B}^*_k$, and the orthogonality condition (1.75) implies $\langle \Psi_\mu, F_N \Psi_0\rangle = 0$ for all $\Psi_\mu \in \mathbb{B}^*_k$. Also, $\langle \Psi_0, F_N \Psi_0\rangle =$ follows from the definition of F_N, so that altogether, (Ψ_0, Λ_0) fulfils

$$\langle \Psi_0, F_N \Psi_\mu\rangle = \Lambda_0 \langle \Psi_0, \Psi_\mu\rangle$$

for all $\Psi_\mu \in \mathbb{B}_k$, and thus solves the weak eigenvalue problem for F_N. To complete the proof, we show that for all $\Psi \in U_0^\perp$,

$$\langle \Psi, F_N \Psi\rangle \geq (\Lambda + \underline{\gamma}) \, \|\Psi\|^2_{\mathbb{L}}, \tag{1.79}$$

where $\overline{\gamma} := \min\{\overline{\Lambda}^* - \overline{\lambda}, \underline{\Lambda}^* - \underline{\lambda}\}$; Λ_0 then is a simple lowest eigenvalue of F_N, and (1.78) follows from Lemma 1.24. First of all, we decompose $\Psi \in U_0^\perp$ into $\Psi = \Psi^1 + \Psi^2$, where Ψ^1 is a linear combination of all Slater determinants which contain at least one function from $\overline{V}^+$, and Ψ^2 thus consists of Slater determinants containing at least one function from $\overline{V}^-$. From the orthogonality of orbitals of different spins and (3.4), it is straightforward to see that

$$\langle F_N\Psi, \Psi\rangle \;=\; \langle F_N\Psi^1, \Psi^1\rangle + \langle F_N\Psi^2, \Psi^2\rangle, \qquad \langle \Psi, \Psi\rangle \;=\; \langle \Psi^1, \Psi^1\rangle + \langle \Psi^2, \Psi^2\rangle.$$

Ψ^1 can be written as $\Psi^1 = \mathcal{Q}(\Phi^+ \otimes \Phi^-)$, where Φ^+ is a linear combination of Slater determinants from the k-fold tensor product space formed only from "spin up" orbitals with indices from $\mathcal{I}^+$, Φ^- is analogously from the $(N-k)$-fold tensor product space formed from orbitals with indices from $\mathcal{I}^-$, and $\mathcal{Q}$ is the antisymmetrization operator, see Section 1.4. Again, orthogonality of different spins and (3.4) applies to give

$$\langle F_N\Psi^1, \Psi^1\rangle \;=\; \langle F_k\Phi^+, \Phi^+\rangle + \langle F_{N-k}\Phi^-, \Phi^-\rangle.$$

We note that F is also symmetric on V^+ and V^- by (1.75), and denote the eigenvalues of F on V^+ and V^- by $\lambda_{\overline{1}} \le \ldots \le \lambda_{\overline{k}}$ and $\lambda_{\underline{k+1}} \le \ldots \le \lambda_{\underline{N}}$, respectively. We now show

$$\langle F_k\Phi^+, \Phi^+\rangle \;\;\ge\;\; \sum_{i=1}^{k-1} \lambda_{\overline{i}} \;+\; \overline{\Lambda}^*, \qquad \langle F_{N-k}\Phi^-, \Phi^-\rangle \;\;\ge\;\; \sum_{i=k+1}^{N} \lambda_{\underline{i}}\,; \tag{1.80}$$

the analogous argument applied to Ψ^2 with interchanged spins then on the whole implies (1.79). To do so, we decompose further, $\Phi^+ = \sum_{r=1}^k \Phi_r$, where Φ_r is a determinant containing exactly r orbitals from $\overline{V}^+$ (and thus $k-r$ from V^+). Again using the orthogonality condition (3.4), it follows that $\langle F_k\Phi^+, \Phi^+\rangle = \sum_{r=1}^N \langle F_k\Phi_r, \Phi_r\rangle$. We fix r and use orthogonality again to get

$$\langle F_k\Phi_r, \Phi_r\rangle \;=\; \langle F_{k-r}\Phi_r^1, \Phi_r^1\rangle + \langle F_r\Phi_r^2, \Phi_r^2\rangle,$$

in which Φ_r^1 is a Slater determinant consisting of $(k-r)$ functions from V^+, while Φ_r^2 contains r functions from $\overline{V}^+$. By expanding each of the one-particle functions contained in Φ_r^1 into a one-particle eigenbasis $\tilde{\chi}_I$ of F on V^+, it is not hard to see that

$$\langle F_{k-r}\Phi_r^1, \Phi_r^1\rangle \;\;\ge\;\; \Lambda_r\langle \Phi_r^1, \Phi_r^1\rangle, \qquad \Lambda_r := \sum_{i=1}^{k-r} \lambda_{\overline{i}}.$$

For Φ_r^2, Lemma 1.25 applies to $F_r|_{\overline{V}^+}$ to give $\langle F\Phi_r^2, \Phi_r^2\rangle \ge r\Lambda^*$. Because this holds for any $r \in \left]k\right]$, this means

$$\langle F\Phi^+, \Phi^+\rangle \;\;\ge\;\; (\Lambda_1 + \Lambda^*)\,\langle \overline{\Phi}, \overline{\Phi}\rangle.$$

A similar argument (in which the only difference is that we have to include $r = 0$) shows

$$\langle F\Phi^-, \Phi^-\rangle \;\;\ge\;\; \sum_{j=k+1}^{N} \lambda_{\underline{i}},$$

which finally yields (1.80), thus completing the proof.

□

1.7 Conclusions - Towards discretisation

In this section, we have shown that the problem of determining the ground state energy of an electronic system is equivalent to computing the $\lfloor N/2 \rfloor + 1$ respective lowest eigenvalues of the weak Schrödinger equations (1.36), i.e.

$$\langle (\hat{H} - E^*)\underline{\Psi},\ \Psi_\mu \rangle \ = \ 0 \quad \text{for all} \ \ \Psi_\mu \in \mathbb{B}_k \tag{1.81}$$

on the spaces $\mathbb{L}^2_k$, $0 \leq k \leq N/2$, where the matrix elements $\langle \hat{H}\Psi_\mu, \Psi_\nu \rangle$ were given in Definition 1.21. Note that when viewed as an "infinite dimensional Galerkin scheme" (i.e a generalized Fourier ansatz), (1.81) yields a system of infinitely many equations for a coefficient vector $(c_\mu)_{\mu \in \mathcal{M}} \in \ell_2(\mathcal{M})$ for which

$$\underline{\Psi} = \sum_{\mu \in \mathcal{M}} c_\mu \Psi_\mu \in \mathbb{H}^1_k$$

solves (1.81), and this set of equations is still equivalent to the original problem of computing the ground state energy. In contrast, common modeling processes in Quantum Chemistry normally start rightaway with a Galerkin discretisation

$$\hat{\mathbf{H}} \ = \ \sum_{P,Q \in \mathcal{I}^{\text{disc}}} h_{P,Q} a^\dagger_P a_Q \ + \ \frac{1}{2} \sum_{P,Q,R,S \in \mathcal{I}^{\text{disc}}} \langle PQ \| RS \rangle a^\dagger_P a^\dagger_Q a_S a_R, \tag{1.82}$$

of the original Hamiltonian $\hat{H}$, in which $\mathcal{I}^{\text{disc}} \subseteq \mathcal{I}$ is a finite selection of indices, and the corresponding tensor basis used is constructed from a finite selection

$$B^{\text{disc}}_k \ = \ \{\varphi_p \in H^1(\mathbb{R}^3) |\ p \in D]\}$$

of D spatial orbitals. In the context of quantum chemistry, the Galerkin method is usually termed "Configuration Interaction method" (CI), and the space spanned by the corresponding functions $\Psi_\mu, \mu \in \mathcal{I}^{\text{disc}}$ is usually termed the "full CI space", in contrast to the continuous, so-called "complete CI space" $\mathbb{H}^1_k$. If, for discretisation, we choose a one-particle-basis B^{disc}_k of spatial orbitals containing at least $D \geq N - k$ elements, the according discretised tensor basis $\mathbb{B}^{disc}_k$ of the discrete tensor subspace of $\mathbb{H}^1_k$ is of the cardinality $\binom{D}{k}\binom{D}{N-k}$. This space is usually much too large for computational practice, and a further reduction of the complexity of the used discrete model is inevitable for practical computations, thus leading to (discrete versions of) the different methods introduced in Sections 2 and 3, approximating the solution of the eigenvalue equation for the discretised Hamiltonian $\hat{\mathbf{H}}$.

It is essential to note though that $\hat{\mathbf{H}}$ does not necessarily reflect the properties of $\hat{H}$; for example, every discrete Hamiltonian $\hat{\mathbf{H}}$ admits a complete eigenbasis (due to its property of being symmetric), but the continuous Hamiltonian does not need to have this property; see Section 1.3. Taking the discrete ansatz (1.82) as starting point for a numerical analysis

is therefore unsatisfactory, because for example, it not *a priori* self-evident that eigenvalues of $\hat{\mathbf{H}}$ approximate eigenvalues of $\hat{H}$ if the basis set size is increased.[21] Additionally, common mathematical concepts like quasi-optimality of discrete solutions or goal-oriented error estimators naturally involve estimates with respect to the real, continuous solution $\underline{\Psi}$ or in terms of the best approximation error $\inf_{\Psi^{\mathrm{disc}} \in \mathbb{H}_k^{1,disc}} \|\Psi^{\mathrm{disc}} - \underline{\Psi}\|$; estimates of this kind are thus *a priori* excluded in the discrete setting, and we will prove some estimates of that same style in Section 3 for the Coupled Cluster method in the continuous setting. The discrete setting also has certain short-comings in view of the construction and analysis of adaptive variants of the methods of quantum chemistry in the vein of [52, 53, 54], as was for instance performed by the author and colleagues for the simple eigenvalue problem including the "complete CI" problem (1.81) in [58, 181]: Usually, the first step in the analysis performed is to formulate the method and iteration scheme under consideration in the original, infinite-dimensional space and to establish convergence results for that scheme; afterwards, the finite dimensional approximation can interpreted as a (controllable) perturbation of the infinite dimensional scheme.[22] Because application of these methods to the minimization framework of Section 2 and to the Coupled Cluster method analysed in Section 3 is desirable, we will prove such convergence results for the continuous versions of the methods of DFT, Hartree-Fock, CI and Coupled Cluster in the next sections.

Motivated by the above reasoning, the methods and algorithms of quantum chemistry to be introduced and analysed in the following sections will, in contrast to what is normally done in the literature, be formulated in the continuous spaces $\mathbb{L}_k^2$ resp. $\mathbb{H}_k^1$, and error estimates will, unless explicitly stated otherwise, apply to both the continuous space and – by replacing the respective space under consideration with a discretisation of it – also to the discretised setting, in which case the estimates are uniform with respect to discretisation parameters.

[21] See [214] for a short analysis of the Galerkin/CI method, relating the quality of discrete eigenvalue and eigenvector approximations to the error of the best approximation in the chosen subspace.

[22] For a brief introduction to this approach in the context of quantum chemistry calculations, see also [76].

2 Analysis of a "direct minimization" algorithm used in Hartree-Fock, DFT and CI calculations

With the results and assumptions of Section 1, computation of the eigenpair (Ψ, E^*) solving Problem 1.12 is equivalent to computing the minimizer and minimum of the functional

$$\mathcal{J}_{\hat{H}} : \ H^1\backslash\{0\} \ \to \ \mathbb{R}, \quad \mathcal{J}_{\hat{H}}(\Psi) = \frac{\langle \hat{H}\Psi, \Psi\rangle}{\langle \Psi, \Psi\rangle}, \tag{2.1}$$

with $\hat{H}$ given by (1.64) and $H^1 := \mathbb{H}^1_k$ for fixed spin number k. The methods of Hartree-Fock (HF), Density Functional Theory (DFT) and Configuration Interaction (CI), being subject of the present Section 2, treat this classical minimization task for the *Rayleigh quotient* (2.1) directly, while they reduce the complexity of this task by restricting the admissible space for the minimizer in one way or another. In all cases, this proceeding leads to a constrained minimization task that then has to be treated with a suitable algorithm.

This section is dedicated to the analysis of a preconditioned steepest descent (or "direct minimization") algorithm popular in the treatment of those minimization tasks, especially in the context of HF and DFT, but also suitable for the CI method or more generally, in the context of invariant subspace computation. Formulated abstractly, the direct minimization algorithm to be formulated below can be used to treat the following minimization task for a suitable energy functional $\mathcal{J}$:

Problem 2.1. *(Minimization under orthogonality constraints)*

For a fixed Gelfand triple $V \hookrightarrow X \hookrightarrow V'$, minimize a given, sufficiently often differentiable functional

$$\mathcal{J} : V^N\backslash\{0\} \to \mathbb{R}, \quad \mathcal{J}(\Phi) \ = \ \mathcal{J}(\varphi_1, \ldots, \varphi_N), \tag{2.2}$$

which is

(a) invariant with respect to unitary transformations, i.e.

$$\mathcal{J}(\Phi) = \mathcal{J}(\Phi\mathbf{U}) = \mathcal{J}\big((\sum_{j=1}^{N} u_{i,j}\varphi_j)_{i=1}^N\big), \tag{2.3}$$

for any orthogonal matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$, and

(b) subordinated to the orthogonality constraints

$$\langle \varphi_i, \varphi_j\rangle \ := \ \int_{\mathbb{R}^3} \varphi_i(x)\varphi_j(x)dx \ = \ \delta_{i,j}. \tag{2.4}$$

We will start out treatment in Section 2.1 by shortly introducing the ansatzes of HF, DFT and CI, and show that they all lead to minimization problem 2.1. We will then re-write this problem as one posed on the infinite-dimensional analogue of the *Grassmann manifold* [10] in Section 2.2. In the subsequent Section 2.3, we introduce the above mentioned direct minimization algorithm corresponding to a quasi-Newton method [60, 163] on $\mathcal{G}$; we then use this theoretical framework to prove local convergence results for the algorithm which generalize analogous results for eigenvalue computations (see the references in Section (2.3)(ii)). Results include linear convergence of the iterates (Theorem 2.14), quadratic dependence of the energies on the error of the eigenfunctions (Theorem 2.15), and residual estimators (Lemma 2.16).

Please note that Sections 2.2, 2.3 reuse (re-edited) parts of the author's own work contributed to [191] under the advisory of R. Schneider. In particular, Lemma 2.5 below has been the work of J. Blauert and is included only for reasons of completeness, but quoted without proof (given in [191]).

2.1 Overview: The Hartree-Fock/Kohn-Sham model and the CI method

(i) The Hartree-Fock and Kohn-Sham model. The first two methods we introduce are the probably most important ones for the qualitative study of larger systems: The classic *ab initio* Hartree-Fock (HF) method [103], analysed for instance in [15, 139, 140, 143, 153], and the semi-empirical Kohn-Sham (KS) method [63, 165, 186] of Density Functional Theory (DFT), introduced in [109, 125, 145] and recently analysed in [7]. Nowadays, these two methods are standardly applied to systems with several hundred atoms, providing energies up to a relative error of about 10^{-2} and other properties within $5-10\%$ accuracy (see e.g. [103]). Present scientific efforts in further practical development of HF and DFT often concentrate on reduction of the canonical N^3 scaling [191] to linear scaling methods, see e.g. [59, 83, 87, 88, 124, 138, 151, 161]. We now at first briefly introduce the basic ideas of the HF method; we will then sketch the modifications that are necessary to obtain the KS method.

The Hartree-Fock method replaces the high-dimensional problem of minimizing (2.1), i.e. that of the computation of a convergent sum $\Psi = \sum_{\mu\in\mathcal{I}} c_\mu \Psi_\mu$ of Slater determinants from $\mathbb{B}_k$ for fixed $0 \le k \le N/2$, by the one of finding a vector of N functions

$$\Phi \;=\; (\varphi_1, \ldots, \varphi_N) \;\in \big(H^1(\mathbb{R}^3, \mathbb{R})\big)^N$$

such that the Slater determinant

$$\Psi_{HF,k}(\varphi_1, \ldots, \varphi_N) \;:=\; \mathcal{Q}\big(\chi_{\overline{1}} \otimes \ldots \otimes \chi_{\overline{k}} \otimes \chi_{\underline{k+1}} \otimes \ldots \otimes \chi_{\underline{N}}\big), \tag{2.5}$$

with the spin orbitals χ_I constructed from the functions φ_i along the lines of Section 1.4 and $\mathcal{Q}$ defined by (1.45), minimizes the functional $\mathcal{J}_{\hat{H}}$ from (2.1) over the set of all such possible antisymmetrized rank-1-tensors of spin number k, i.e. $(\varphi_1, \dots, \varphi_N)$ is the minimizer of the (unrestricted) HF functional [174][23,24]

$$\mathcal{J}_{HF,k} : \quad \big(H^1(\mathbb{R}^3)\backslash\{0\}\big)^N \;\to\; \mathbb{R}, \quad (\varphi_1, \dots, \varphi_N) \;\mapsto\; \mathcal{J}_{\hat{H}}\big(\Psi_{HF,k}(\varphi_1, \dots, \varphi_N)\big). \tag{2.6}$$

Computational costs are usually reduced further by use of the *Restricted Open Shell Hartree-Fock model* (ROHF), where the k spatial orbitals φ_i used to construct the "spin up" functions $\chi_{\bar{i}}$ are also used to construct the "spin down" functions $\chi_{\underline{k+i}}, i = 1, \dots, k$, so that only $N - 2k$ additional "spin down" functions $\varphi_i, i = 2k+1, \dots, N$ and on the whole $N - k$ functions have to be computed, see e.g. [103, 201].
If for a molecule with an even number $N = 2N^*$ of electrons, one assumes z-spin zero (i.e. $k = N/2$, see Section 1.2), this results in the *Closed Shell Restricted Hartree-Fock model* (RHF), which replaces (2.6) by a spin-free model for $N^* = N/2$ *pairs* of electrons, so that

$$\Phi = (\varphi_i)_{i=1}^{N^*} \in H^1(\mathbb{R}^3)^{N^*}.$$

Abbreviating

$$V(x) := -\sum_{\nu=1}^{M} \frac{Z_\nu}{\|x - R_\nu\|},$$

the corresponding functional then reads

$$\begin{aligned}\mathcal{J}_{HF}(\Phi) := \sum_{i=1}^{N^*} \int_{\mathbb{R}^3} \Big(& \frac{1}{2}|\nabla\varphi_i(x)|^2 + V(x)|\varphi_i(x)|^2 \;+\; \sum_{j=1}^{N^*} \int_{\mathbb{R}^3} \frac{|\varphi_j(y)|^2}{\|x-y\|}\, dy\, |\varphi_i(x)|^2 \\ & - \frac{1}{2}\sum_{j=1}^{N^*} \int_{\mathbb{R}^3} \frac{\varphi_i(x)\varphi_j(x)\varphi_j(y)\varphi_i(y)}{\|x-y\|}\, dy\Big)\, dx. \end{aligned} \tag{2.7}$$

We will in the following use the above closed-shell functional $\mathcal{J}_{HF}$ and the corresponding functional $\mathcal{J}_{KS}$ (see below) as prototypes to exemplify the structure of the Hartree-Fock/Kohn-Sham method, e.g. for the derivation of the Fock operator from $\mathcal{J}_{HF}$.
From Lemma 1.17, it is not hard to see that the admissible set for the RHF and ROHF problem can be reduced to $N - 2k$ mutually orthonormal functions $\varphi_i \in H^1(\mathbb{R}^3)$ without changing the minimum, giving the benefit that the evaluation of $\mathcal{J}_{HF}$ can now be

[23] Although minimization of the functional given here is usually called the "unrestricted HF ansatz", this nomenclature does not seem to be unambiguous in the literature. In [15, 139, 143], for instance, minimization of the functional $\mathcal{J} : \big(H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})\backslash\{0\}\big)^N \to \mathbb{R}, \Phi = (\chi_1, \dots, \chi_N) \;\mapsto\; \mathcal{J}_{\hat{H}}(\widehat{\otimes}_{i=1}^N \chi_i)$, i.e. an ansatz without fixed spin number k, is also termed "unrestricted HF ansatz". Note though that this ansatz does not necessarily yield eigenfunctions of the z-spin operator any more.

[24] Note that solutions of the discretised problem now are of size $D \cdot N$ instead of D^N for (2.1). This ansatz is an example for the basic idea of low-rank tensor approximation, and the functional $\mathcal{J}_{\hat{H}}(\mathcal{Q}\,\cdot\,)$ fits into the more general framework of tensor minimization treated by the author and colleagues in [70].

performed with the aid of the Slater-Condon rules, see [201]. Therefore, (2.6) is usually treated as a constrained minimization problem under the condition that

$$\langle \varphi_i, \varphi_j \rangle = \delta_{i,j} \quad \text{for} \quad i,j \in N] \,.$$

Additionally, $\mathcal{J}(\Phi) = \mathcal{J}(\Phi \mathbf{U})$ holds for any unitary $\mathbf{U}$ [103]; thus, the above RHF and ROHF functionals are the first examples for a minimization task of the type of Problem 2.1, where

$$V = H^1(\mathbb{R}^3, \mathbb{R}), \quad X = L^2(\mathbb{R}^3, \mathbb{R})$$

or $V = H^1(\mathbb{R}^3 \times \Sigma^1, \mathbb{R}), X = L^2(\mathbb{R}^3 \times \Sigma^1, \mathbb{R})$ for other variants including a spin variable.

The energy functional of the *Kohn-Sham (KS) model* of DFT, sharing the properties (a) and (b) with the Hartree-Fock functional, can be derived from the Hartree-Fock energy functional by replacing the nonlocal and therefore computationally costly exchange term in the Hartree-Fock functional (i.e. the fourth term in (2.7) in the closed-shell case) by an additional (a priori unknown) exchange correlation energy term $E_{xc}(n)$ depending only on the electron density,[25] given by

$$n(x) = \sum_{i=1}^{N} |\varphi_i(x)|^2.$$

The resulting energy functional for a vector $\Phi = (\varphi_i)_{i=1}^{N^*}$ of orthonormal functions reads

$$\mathcal{J}_{KS}(\Phi) = \frac{1}{2} \sum_{i=1}^{N} \int\limits_{\mathbb{R}^3} |\nabla \varphi_i(x)|^2 dx + \int\limits_{\mathbb{R}^3} n(x) V(x) + \int\limits_{\mathbb{R}^3} \int\limits_{\mathbb{R}^3} \frac{n(x) n(y)}{\|x - y\|} \, dxdy + E_{xc}(n).$$

Because $E_{xc}(n)$ and its analytical properties are unknown, it has to be approximated, e.g. by the simple local density approximation (LDA). We will not touch deeper on this subject here, but refer the reader to the monographs [63, 165]. We only note concludingly that a combination of both HF and DFT models, namely the hybrid B3LYP, is experienced to provide the best results in benchmark computations.

Minimizers of Hartree-Fock or Kohn-Sham type functionals are named Hartree-Fock or Kohn-Sham ground states, respectively. While the question of uniqueness of such ground states is an open problem, their existence has been proven for some Hartree-Fock type functionals in the case that $\sum_{\mu=1}^{K} Z_\mu \geq N$ [139, 143], and recently under the same condition for Kohn-Sham type functionals [7]. For extremely negatively charged molecules, it can be shown that there cannot exist a HF ground state [140]. Minimizers of $\mathcal{J}_{HF,k}$ enjoy similar properties as the ones discussed for the eigenfunctions of $\hat{H}$ in Section 1.3(v), also cf. [139] for details.

[25] This ansatz is justified by one of the cornerstones of DFT, the Hohenberg-Kohn theorem, cf.[109].

(ii) Post-Hartree-Fock methods: General introduction. If one is interested in certain quantities as energy differences or spectral properties, an accuracy of the energies of about $10^{-3} - 10^{-4} E_h$ (the so-called chemical accuracy, corresponding to relative accuracies of about 10^{-5} to 10^{-7}) up to $10^{-6} E_h$ (spectroscopic accuracy) is needed in order to compete with practical experiments. In HF and KS methods, effects of electronic correlation are neglected or only treated approximately [103, 153]; in particular, the electron-electron-cusps discussed in Section 1.3(v) are not approximated well, which usually results in a relative modeling error much bigger than the one required to achieve chemical accuracy.
Therefore, so-called *Post-Hartree-Fock methods*, of which the Configuration Interaction (CI) and the Coupled Cluster (CC) method are the most notable iterative examples, take the antisymmetrized rank-1-approximation $\Psi_0 = \Psi_{HF,k}$ provided by a preliminary UHF or ROHF calculation as a starting point for the solution of an appropriate discretisation of the original weak eigenvalue problem (1.36) posed in the tensor space $H^1 = \mathbb{H}^1_k$. A discrete orthonormal basis $B^{disc} = \{\varphi_1, \ldots, \varphi_D\}$ needed for building up subspaces of H^1 and for the evaluation of the matrix elements of the Hamiltonian (see Def. 1.21) is then usually provided by the eigenfunctions of the (discretised) *Fock operator* $\mathbf{F} = \mathbf{F}_{\Phi_0}$ (see Remark 2.7) belonging to the minimizer of the corresponding HF functional. In particular, the eigenfunctions from B^{disc} belong to the lowest eigenvalues form the Hartree-Fock solution $\Psi_0 \in H^1$. This choice for B has the positive effect that the Fock operator $\mathbf{F}_{\Phi_0}$ can be "lifted" to a diagonal Fock operator $\mathbf{F}_{\Phi_0,N}$ defined on the tensor space H^1, usually giving an efficient preconditioner for the solvers used in post-HF calculations, see e.g [103, 201] and also Remark 3.14. In many cases, the reference guess Ψ_0 is sufficiently close to the sought solution $\underline{\Psi}$ of the original eigenvalue equation, so that the mostly Newton-like methods used as solvers for the resulting equations then converge to $\underline{\Psi}$. However, cases are known where convergence is slow or fails, and multi-reference techniques have to be chosen as initial guess to overcome this weakness, cf. e.g. the discussion in [103].

(iii) The Configuration Interaction method, invariant subspace computations. It was already mentioned that CI corresponds to a direct Galerkin discretisation for the weak eigenvalue problem for the Hamiltonian formulated in Section 1.5. For the reasons discussed at the end of Section 1, the tensor basis generated by $B = \{\varphi_1, \ldots, \varphi_D\}$ contains for all but very small molecules far too many basis functions for a numerical computation. Thus, selection rules have to be devised for the set $\mathbb{B}^{\mathrm{disc}}_k$ of those basis functions from $\mathbb{B}_k$ that are included in the calculation, resulting in the discretised minimization task of computing the minimizer and minimum of the restriction of $\mathcal{J}_{\hat{H}}$ to the subspace spanned by $\mathbb{B}^{\mathrm{disc}}_k$,

$$\mathcal{J}_{\hat{\mathbf{H}}} := \mathcal{J}_{\hat{H}}|_{\mathrm{span}\mathbb{B}^{\mathrm{disc}}_k}. \tag{2.8}$$

The canonical way of proceeding is to choose basis functions according to simple selection rules based on excitation levels, e.g. only those functions from the tensor basis $\mathbb{B}_k$ are

included which differ from Ψ in at most two one particle spin basis functions.[26] More sophisticated methods are based on screening procedures using perturbative arguments, as the canonical orbital based methods like CIPSI [71, 101, 112] or [39, 40], and selective Multi-Reference-CI methods [37, 69, 99], or use locality criteria together with localized basis functions [149, 150].
The significance of the CI method for "real life" applications is limited due to lack of size consistency [160] for any truncated basis set. Nevertheless, so-called "Full CI" calculations using the full tensor basis for benchmark computations for small molecules are an important tool when analysing the basis set error induced by the basis in which the HF solutions are computed. Also, from a mathematical viewpoint, CI has the advantage that the relations determining the analytic properties of the Galerkin method are clear to a broad extent [47, 214]; they may therefore be used to deduce analogous analytic results on other post-HF methods, see e.g. [190]. Moreover, modifications of the CI method, e.g. in the context of an adaptive treatment of the equations (see [58, 181] and the remarks at the end of Section 1), may serve as a prototype for an analysis of modifications of other quantum chemical methods in the sense that properties that can be shown to hold for the CI methods may then be transferred to other, more advanced methods as, for instance, the Coupled Cluster method.
As a weak operator eigenvalue problem for the smallest eigenvalue of $\hat{H}$, CI is the special case $N = 1$ of the problem of computing N eigenvalues that form the bottom of the spectrum of a symmetric operator A. In Lemma 2.9, it will be shown that this problem is equivalent to the following Problem 2.2, also covering in a more general context the calculation of an invariant subspace of a symmetric operator A, spanned by the eigenvectors belonging to the lowest N eigenvalues.

Problem 2.2. *(Eigenvalue problem/invariant subspace calculation)*
For a symmetric operator $A : V \to V'$, minimize

$$\mathcal{J}_A(\varphi_1, \ldots, \varphi_N) := \sum_{i=1}^{N} \langle \varphi_i, A\varphi_i \rangle \tag{2.9}$$

among all those $(\varphi_1, \ldots, \varphi_N)$ for which $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$.

□

It is not hard to see that $\mathcal{J}_A$ also fulfils property (a) and that thus Problem 2.1 covers the minimization task associated with CI (where $V = \mathbb{H}^1_k, X = \mathbb{L}_k$) and also that of computing an invariant subspace. For $N > 1$ and $A = \hat{H}$, Problem 2.2 represents the problem of simultaneous computation of E^* and the next $N - 1$ greater energy eigenvalues of $\hat{H}$, which might be interesting if one is interested in excitation energies or opto-electronical properties of the molecule.

[26] The quantum chemical terminology for this is that "only Single and Double excitations of Ψ_0 are included", resulting in the abbreviation CISD for this well-known discretisation.

2.2 Minimization problems on Grassmann manifolds

In Problem 2.1, the invariance (a) of the functional $\mathcal{J}$ with respect to uniform transformations among the eigenfunctions shows a certain redundancy inherent in its formulation. We will now factor out this redundancy; together with the orthogonality constraint (b), this results in the *Grassmann manifold* $\mathcal{G}$, originally defined in finite dimensional Euclidean Hilbert spaces [10], see also [2] for an extensive exposition. We generalize this concept to the present infinite dimensional space V^N and use it to re-state Problem 2.1 as a minimization problem for the according functional $\mathcal{J}$ defined on $\mathcal{G}$ in part (ii) of this section; in part (iii), we will formulate optimality conditions for the treatment of the minimization problem 2.1. In part (iv), those optimality criteria are specified more explicitly for the concrete applications of quantum chemistry introduced in the last section.

(i) Basic notations. For some measurable set Ω, we let $X := L^2(\Omega, \mathbb{R}), X := L^2(\Omega, \mathbb{C})$, or a closed subspace of that spaces. We will work with a Gelfand triple $V \subset X \subset V'$ with the usual L^2-inner product $\langle \cdot, \cdot \rangle$ as dual pairing on $V' \times V$, where either $V := H^t(\Omega)$, $t \geq 0$, see Section 1.1(iii), or an appropriate subspace, for instance corresponding to a Galerkin discretisation. The optimization problem will be formulated on an admissible subset of V^N below, and we prepare this by extending inner products and operators from V to V^N by the following definition.

Definition 2.3. (Inner products, operators and operations on V^N)

For $\Phi = (\varphi_1, \ldots, \varphi_N) \in V^N, \Phi' = (\psi_1, \ldots, \psi_N) \in (V^N)' = (V')^N$, and the L^2-inner product $\langle \cdot, \cdot \rangle$ given on $X = L^2$, we denote

$$\left\langle \Phi^T \Phi' \right\rangle := (\langle \varphi_i, \psi_j \rangle)_{i,j=1}^N \in \mathbb{R}^{N \times N} ,$$

and introduce the dual pairing

$$\langle\langle \Phi', \Phi \rangle\rangle := \operatorname{tr} \left\langle \Phi^T \Phi' \right\rangle = \sum_{i=1}^N \langle \varphi_i, \psi_i \rangle$$

on $(V')^N \times V^N$. Because there holds $V^N = V \otimes \mathbb{R}^N$, we can canonically expand any operator $R : V \to V'$ to an operator

$$\mathcal{R} := R \otimes I : \quad V^N \ \to \ (V')^N, \quad \Phi \ \mapsto \ \mathcal{R}\Phi \ = \ (R\varphi_1, \ldots, R\varphi_N). \tag{2.10}$$

Throughout this section, for an operator $V \to V'$ denoted by a capital letter as $A, B, D, \ldots$, the same calligraphic letter $\mathcal{A}, \mathcal{B}, \mathcal{D}, \ldots$ will denote this expansion to V^N.
Further, we will make use of the following operations: For $\Phi \in V^N$ and $\mathbf{M} \in \mathbb{R}^{N \times N}$, we define the vector $\Phi\mathbf{M} = (I \otimes \mathbf{M})\Phi \in V^N$ by

$$(\Phi\mathbf{M})_j := \sum_{i=1}^N m_{i,j} \varphi_i,$$

cf. also the notation in (2.3), and for $\varphi \in V$ and $v = (v_1, \dots, v_N) \in \mathbb{R}^N$ the element $\varphi \otimes v \in V^N$ by $(v_1\varphi, \dots, v_N\varphi)$.
Finally, we denote by $O(N)$ the orthogonal group of $\mathbb{R}^{N\times N}$.

(ii) The geometry of Stiefel and Grassmann manifolds. Let us now introduce the admissible manifold and prove some of its basic properties. Note in this context that well established results of [10] for the case in the finite dimensional Euclidean spaces cannot be applied to our setting without further difficulties, because the norm induced by the L^2- inner product is weaker than the norm on $V = H^t(\Omega)$.

Our aim is to minimize $\mathcal{J}(\Phi)$ under the orthogonality constraint $\langle \varphi_i, \varphi_j \rangle = \delta_{i,j}$, i.e.

$$\left\langle \Phi^T \Phi \right\rangle = \mathbf{I} \in \mathbb{R}^{N\times N}. \tag{2.11}$$

The subset of all $\Phi \in V^N$ satisfying the property (2.11) is the *Stiefel manifold* [10]

$$\mathcal{V}_{V,N} := \{\Phi = (\varphi_i)_{i=1}^N \mid \varphi_i \in V, \quad \left\langle \Phi^T \Phi \right\rangle - \mathbf{I} \;=\; \mathbf{0} \in \mathbb{R}^{N\times N}\},$$

i.e. the set of all orthonormal bases of N-dimensional subspaces of V.

All functionals $\mathcal{J}$ under consideration are unitarily invariant, i.e. there holds (2.3). To abolish this nonuniqueness, we will identify all orthonormal bases $\Phi \in \mathcal{V}_{V,N}$ spanning the same subspace $V_\Phi := \text{span}\,\{\varphi_i : i = 1, \dots, N\}$. To this end we consider the *Grassmann manifold*, defined as the quotient

$$\mathcal{G}_{V,N} := \mathcal{V}_{V,N}/\!\sim$$

of the Stiefel manifold with respect to the equivalence relation $\Phi\sim\widetilde{\Phi}$ if $\widetilde{\Phi} = \Phi\mathbf{U}$ for some $\mathbf{U} \in O(N)$. We usually omit the indices and write $\mathcal{V}$ for $\mathcal{V}_{V,N}$, $\mathcal{G}$ for $\mathcal{G}_{V,N}$ respectively. To simplify notations we will often also work with representatives instead of equivalence classes $[\Phi] \in \mathcal{G}$.
The interpretation of the Grassmann manifold as equivalence classes of orthonormal bases spanning the same N-dimensional subspace is just one way to define the Grassmann manifold. We can as well identify the subspaces with orthogonal projectors onto these spaces. To this end, let us for $\Phi = (\varphi_1, \dots, \varphi_N) \in \mathcal{V}$ denote by D_Φ the L^2-orthogonal projector onto $\text{span}\{\varphi_1, \dots, \varphi_N\}$. One straight-forwardly verifies

Remark 2.4. *There is a one-to-one relation identifying $\mathcal{G}$ with the set of rank-N L^2-orthogonal projection operators D_Φ.*

The following well-known representation of the tangent space of the Grassmann manifold will be needed later. See [98] or [191] for the proof.

Lemma 2.5. *(Tangent space of $\mathcal{G}$)*
The tangent space of the Grassmann manifold $\mathcal{G}$ at $[\Phi] \in \mathcal{V}$ is

$$\begin{aligned} \mathcal{T}_{[\Phi]}\mathcal{G} &= \{W \in V^N | \left\langle W^T \Phi \right\rangle = \mathbf{0} \in \mathbb{R}^{N\times N}\} \\ &= (\mathrm{span}\{\varphi_1, \ldots, \varphi_N\}^\perp)^N . \end{aligned}$$

Thus, the operator $(\mathcal{I} - \mathcal{D}_\Phi)$, where D_Φ is the L^2-projector onto the space spanned by Φ and $\mathcal{D}_\Phi$ is its expansion as above, is an L^2-orthogonal projection from V^N onto the tangent space $\mathcal{T}_{[\Phi]}\mathcal{G}$.

To end this section, we prove a geometric result needed later.

Lemma 2.6. *(Differences and projected differences)*
Let $[\Phi_0] \in \mathcal{G}$, $D = D_{\Phi_0}$ be the L^2-projector on $\mathrm{span}[\Phi_0]$ *and $\|.\|$ the norm induced by the L^2 or H^1 inner product. For any orthonormal set $\Phi = (\varphi_1, \ldots, \varphi_N) \in \mathcal{V}$ sufficiently close to $[\Phi_0] \in \mathcal{G}$ in the sense that for all $i \in \{1, \ldots, N\}$, $\|(I - D)\varphi_i\| < \delta$, there exists an orthonormal basis $\bar{\Phi}_0 \in \mathcal{V}$ of* $\mathrm{span}[\Phi_0]$ *for which*

$$\Phi - \bar{\Phi}_0 \quad = \quad (I - \mathcal{D})\Phi \; + \; \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2).$$

Proof. For $i = 1, \ldots, N$, let

$$\psi_i' \quad = \quad \arg\min\{\|\psi - \varphi_i\|, \psi \in \mathrm{span}\{\psi_i | i = 1, \ldots, N\}, \|\psi\| = 1\} \quad = \quad D\varphi_i / \|D\varphi_i\|,$$

and set $\Phi_0' := (\psi_1', \ldots, \psi_N')$. If we denote by P_i' the L^2 projector on the space spanned by ψ_i', it is straightforward to see from the series expansion of the cosine that

$$(I - D)\varphi_i \quad = \quad (I - P_i')\varphi_i \quad = \quad \varphi_i - \psi_i' \quad + \quad \mathcal{O}(\|(I - D)\varphi_i\|^2) \tag{2.12}$$

The fact that $\Phi_0' \notin \mathcal{V}$ is remedied by orthonormalization of Φ_0' by the Gram-Schmidt procedure. For the inner products occurring in the orthogonalization process (for which $i \neq j$), there holds

$$\begin{aligned} \langle \psi_i', \psi_j' \rangle &= \langle \psi_i' - \varphi_i, \psi_j' \rangle + \langle \varphi_i, \psi_j' - \varphi_j \rangle + \langle \varphi_i, \varphi_j \rangle \\ &= -\langle (I - D)\varphi_i, \psi_j{}' \rangle - \langle (I - D)\varphi_i, (I - D)\varphi_j \rangle \; + \; \mathcal{O}(\|(I - D)\varphi_i\|^2). \\ &= \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2) \end{aligned}$$

where we have twice replaced $\varphi_i - \psi_i'$ by $(I - D^*)\varphi_i$ according to (2.12) and made use of the orthogonality of D^*. In particular, for Φ sufficiently close to $[\Phi_0]$, the Gramian matrix is non-singular because the diagonal elements converge quadratically to one while the off-diagonal elements converge quadratically to zero. By an easy induction for the orthogonalization process and a Taylor expansion for the normalization process, we obtain that Φ_0' differs from the orthonormalized set $\bar{\Phi}_0 := (\bar{\psi}_1, \ldots, \bar{\psi}_N)$ only by an error term depending on $\|(I - \mathcal{D})\Phi\|^2$. Therefore,

$$\varphi_i - \bar{\psi}_i \quad = \quad \varphi_i - \psi_i' \; + \; \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2) \quad = \quad (I - D)\varphi_i \; + \; \mathcal{O}(\|(I - \mathcal{D})\Phi\|^2),$$

so that

$$\Phi - \bar{\Phi}_0 \;\; = \;\; (I - \mathcal{D})\Phi \;+\; \mathcal{O}(\|(I-\mathcal{D})\Phi\|^2),$$

and the result is proven.

□

(iii) General optimality conditions on the Grassmann manifold. By the first order optimality condition for minimization tasks, a minimizer $[\Phi_0] = [(\psi_1, \ldots, \psi_N)] \in \mathcal{G}$ of the functional $\mathcal{J} : \mathcal{G} \to \mathbb{R}, [\Phi] \mapsto \mathcal{J}(\Phi)$ over the Grassmann manifold $\mathcal{G}$ satisfies

$$\langle\langle \mathcal{J}'(\Phi_0), \delta\Phi \rangle\rangle = 0 \quad \text{for all} \;\; \delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G} \;, \tag{2.13}$$

i.e. the gradient $\mathcal{J}'(\Phi_0) \in (V')^N = (V^N)'$ vanishes on the tangent space $\mathcal{T}_{\Phi_0}\mathcal{G}$ of the Grassmann manifold. This property can also be formulated by

$$\left\langle (\delta\Phi)^T \mathcal{J}'(\Phi_0) \right\rangle = \mathbf{0} \quad \text{for all} \;\; \delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G},$$

or equivalently, by Lemma 2.5 as

$$\langle (\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi_0), \Phi \rangle = 0 \quad \text{for all} \;\; \Phi \in V^N. \tag{2.14}$$

Let the bracket $[\cdot, \cdot\cdot]$ denote the usual commutator, and with $(\mathcal{J}'(\Phi_0))_i \in V'$ the i-th component of $\mathcal{J}'(\Phi_0)$, let $\Lambda = (\langle (\mathcal{J}'(\Phi_0))_j, \psi_i \rangle)_{i,j=1}^N$. In strong formulation, this condition can then be formulated in the various ways

$$(\mathcal{I} - \mathcal{D})\mathcal{J}'(\Phi_0) \;\; = \;\; [\mathcal{J}', \mathcal{D}]\Phi_0 \;\; = \;\; \mathcal{J}'(\Phi_0) - \Phi_0\Lambda \;\stackrel{!}{=}\; 0 \in (V')^N. \tag{2.15}$$

Note that this corresponds to one of the optimality conditions for the Lagrangian yielded from the common approach of the Euler-Lagrange minimization formalism: Introducing the Lagrangian

$$\mathcal{L}(\Phi, \Lambda) := \frac{1}{2}\left(\mathcal{J}(\Phi) + \sum_{i,j=1}^{N} \lambda_{i,j}(\langle \varphi_i, \varphi_j \rangle_{L_2} - \delta_{i,j}) \right), \tag{2.16}$$

and denoting by $\mathcal{L}^{(1,\Phi)}(\Phi, \Lambda)$ the derivative restricted to V^N, the first order condition is then given by

$$\mathcal{L}^{(1,\Phi_0)}(\Phi_0, \Lambda) \;\; = \;\; \mathcal{J}'(\Phi_0) - (\sum_{k=1}^{N} \lambda_{i,k}\psi_k)_{i=1}^N = 0 \in (V')^N. \tag{2.17}$$

Testing this equation with $\psi_j, j = 1, \ldots, N$, verifies that the Lagrange multipliers indeed agree with the Λ defined above, so that (2.14) and (2.17) are equivalent. Note also that the remaining optimality conditions

$$\frac{\partial \mathcal{L}}{\partial \lambda_{i,j}} \;\; = \;\; \frac{1}{2}\left(\langle \psi_i, \psi_j \rangle_{L_2} - \delta_{i,j} \right) \;\; = \;\; 0$$

of the Lagrange formalism are now incorporated in the framework of the Stiefel manifold. Let us denote by $\mathcal{L}^{(2,\Phi)}(\Phi,\Lambda)$ the second derivative of $\mathcal{L}$ with respect to Φ. From the representation (2.15), it then follows that $\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)$, taken at the minimizer Φ_0, is given by

$$\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)\Phi \;=\; \mathcal{J}''(\Phi_0)\Phi - \Phi\Lambda. \tag{2.18}$$

As a necessary second order condition for a minimum, $\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)$ has to be positive semidefinite on $\mathcal{T}_{[\Phi_0]}\mathcal{G}$. For our convergence analysis, we will have to impose the stronger condition on $\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)$ of being elliptic on the tangent space, i.e.

$$\langle\langle \mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)\delta\Phi \;,\; \delta\Phi\rangle\rangle \;\;\geq\;\; \gamma\,\|\delta\Phi\|^2_{V^N}, \qquad \text{for all} \quad \delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G} \tag{2.19}$$

holds for some $\gamma > 0$. It is an unsolved question whether there are general conditions on the functional $\mathcal{J}$ under which (2.19) holds for minimization problems of the type of Problem 2.1, cf. also the remarks in the next section.

(iv) First and second order conditions for problems from quantum chemistry. For the functionals used in the context of HF, DFT and CI calculations, we now take a more explicit look at the first and second order conditions for the functionals. For the explicit derivation of the results in the following remark, see [165, 201].

Remark 2.7. (Fock-/Kohn-Sham operator)
For the functional $\mathcal{J}_{HF}$ of ROHF and RHF, $\mathcal{J}'_{HF}(\Phi) = \mathcal{A}_\Phi\Phi \in (V')^N$, where $A_\Phi = F_{HF,\Phi} : H^1(\mathbb{R}^3) \to H^{-1}(\mathbb{R}^3)$ is the so-called Fock operator and $\mathcal{A}_\Phi$ is defined by A_Φ through (2.10); using the notation of the *density matrix* $\rho_\Phi(x,y) := \sum_{i=1}^N \varphi_i(x)\varphi_i(y)$ and the electron density $n_\Phi(x) := \rho_\Phi(x,x)$ already introduced above, it is in the closed-shell case given by

$$F_{HF,\Phi}\varphi(x) \;:=\; -\frac{1}{2}\Delta\varphi(x) \;+\; V(x)\varphi(x) \;+\; 2\int\limits_{\mathbb{R}^3} \frac{n_\Phi(y)}{\|x-y\|}\,dy\,\varphi(x) \;-\; \int\limits_{\mathbb{R}^3} \frac{\rho_\Phi(x,y)\varphi(y)}{\|x-y\|}\,dy.$$

For the gradient of the Kohn-Sham functional $\mathcal{J}_{KS}$, there holds the following: Assuming that E_{xc} in $\mathcal{J}_{KS}$ is differentiable and denoting by v_{xc} the derivation of E_{xc} with respect to the density n, we have $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi\Phi \in (V')^N$, with $A_\Phi = F_{KS,n}$ the Kohn-Sham Hamiltonian, given in the closed-shell case by

$$F_{KS,n}\varphi(x) \;:=\; -\frac{1}{2}\Delta\varphi(x) \;+\; V(x)\varphi(x) \;+\; 2\int\limits_{\mathbb{R}^3} \frac{n_\Phi(y)}{\|x-y\|}\,dy\,\varphi(x) \;+\; v_{xc}(n(x))\varphi(x).$$

For both functionals, the Lagrange multiplier Λ of (2.17) at a minimizer $\Phi_0 = (\psi_1,\ldots,\psi_N)$ is given by

$$\lambda_{i,j} = \langle A_{\Phi_0}\psi_i, \psi_j\rangle. \tag{2.20}$$

There exists a unitary transformation $\mathbf{U} \in O(N)$ amongst the functions ψ_i, $i \in N]$ such that the Lagrange multiplier is diagonal for $\Phi_0 \mathbf{U} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_N)$,

$$\lambda_{i,j} := \langle A\tilde{\psi}_i, \tilde{\psi}_j \rangle = \lambda_i \delta_{i,j}.$$

so that the ground state of the HS resp. KS functional (i.e. minimizer of $\mathcal{J}$) satisfies the nonlinear *Hartree-Fock* resp. *Kohn-Sham* eigenvalue equations

$$F_{HF,\Phi_0}\psi_i = \lambda_i \psi_i, \quad \text{resp.} \quad F_{KS,n}\psi_i = \lambda_i \psi_i, \qquad \lambda_i \in \mathbb{R}, \qquad i \in N] \,, \tag{2.21}$$

for some $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ and a corresponding set of orthonormalized functions $\Phi_0 = (\psi_i)_{i=1}^N$ up to a unitary transformation $\mathbf{U}$.

□

A result concerning the converse, i.e. if for every collection $\Phi = (\varphi_1, \ldots, \varphi_N)$ belonging to the N lowest eigenvalues of the Fock-/KS operator, the corresponding Slater determinant (2.5) actually gives the Hartree-Fock/DFT energy by $\mathcal{J}_{\hat{H}}(\Psi_{HF,k})$, is unknown. Also, concerning the strengthened second order condition (2.19), it is not clear whether (2.19) holds under certain conditions for the functionals of Hartree-Fock and density functional theory. In the case of Hartree-Fock, it is known that it suffices to demand that $\mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda) > 0$ on $\mathcal{T}_{[\Phi_0]}\mathcal{G}$ because this already implies $\mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda)$ is bounded away from zero, cf. [146].

Remark 2.8. (Upper and lower bound for Fock-/Kohn-Sham operators.)

For later purposes, we note that analogously to the result (1.39) for the Hamiltonian, there holds for the Fock operator $F_{HF,\Phi}$ belonging to a set of functions $\Phi \in \mathcal{V}$ that

$$c \, \|\varphi\|_1^2 - \mu \langle \varphi, \varphi \rangle \quad \leq \quad \langle F_\Phi \varphi, \varphi \rangle \quad \leq \quad C \, \|\varphi\|_1^2 \tag{2.22}$$

for all $\varphi \in H^1(\mathbb{R}^3)$ and some constants $c, C > 0$, $\mu \in \mathbb{R}$.[27] The same result holds if $F_{KS,n}$ is a Kohn-Sham operator in which the exchange term $v_{xc}(n)$ maps $H^1(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$ boundedly (as multiplication operator).[28] Therefore, such operators F can be shifted to elliptic mappings $F + \mu I : H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \to H^{-1}(\mathbb{R}^3\{\pm\frac{1}{2}\})$, cf. Lemma 1.23. In particular, using Lemma 1.25, $F + \mu I$ induces a norm on the tensor space $\mathbb{L}^2$ that is equivalent to the $\mathbb{H}^1$-norm, a useful fact in the theoretical analysis of Post-HF-methods, see e.g. Section 3. For practical issues like preconditioning, we note that the lifted Fock-/KS-operator, shifted by the sum of the lowest N eigenvalues, is under a certain gap condition elliptic on the orthogonal complement of the sought eigenspace, see Lemma 1.26, Remark 3.14.

[27] The proof uses the Hardy inequality [207, 214] and is essentially the same as for the Hamiltonian $\hat{H}$ given in [214], so it is omitted.

[28] Again, cf. the analogous argument from the proof for the Hamiltonian $\hat{H}$ from [214].

For the simplified Problem 2.2, the minimization of $\mathcal{J}_A$ is related to finding an orthonormal basis $\{\psi_i : 1 \leq i \leq N\}$ spanning the invariant subspace of A given by the first eigenfunctions of A, and gives an explicit condition for the uniqueness of the minimizer.

Lemma 2.9. *(Problem 2.2 and invariant subspace calculation)*
Let $A : V \to V'$ be a bounded symmetric operator, fulfilling the Gårding inequality (1.66); denote its spectrum by $\mathrm{spec}(A)$. *The gradient of the functional $\mathcal{J}_A$ from (2.9) belonging to the simplified problem is given by*

$$\mathcal{J}'(\Phi) = \mathcal{A}\Phi \in (V')^N.$$

Φ_0 therefore is a stationary point of the associated Lagrangian $\mathcal{L}$ if and only if there exists an orthogonal transformation $\mathbf{U}$ such that

$$\Phi_0\mathbf{U} = (\tilde{\psi}_1, \ldots, \tilde{\psi}_N) \in V^N$$

consists of N pairwise orthonormal eigenfunctions of A, i.e. $A\psi_i = \lambda_i\psi_i$ for $i = 1, \ldots, N$; in this case, there holds

$$\mathcal{J}(\Phi_0) = \sum_{i=1}^{N} \lambda_i.$$

If A has N lowest eigenvalues $\lambda_1 \leq \ldots \leq \lambda_N$ satisfying the gap condition

$$\lambda_N \quad < \quad \overline{\lambda} \;:=\; \inf\{\lambda \mid \lambda \in \mathrm{spec}(A)\backslash\{\lambda_1, \ldots, \lambda_N\}\}, \tag{2.23}$$

the minimum of $\mathcal{J}_A$ is attained if and only if the corresponding eigenvalues λ_i, $i = 1, \ldots, N$ are the N lowest eigenvalues, and it is unique up to orthogonal transformations.

□

Remark 2.10. Lemma 2.9 implies in particular that under Assumption 1.38, the CI method is well-posed as a minimization method for the Rayleigh quotient (2.1); more generally, the simultaneous computation of N eigenvalues by minimizing (2.9) with $A = H$ is well-posed as long as H has N eigenvalues below the essential spectrum.

Proof of Lemma 2.9. By (2.15), the first order condition for a stationary point implies $\mathcal{A}\Phi_0 = \Phi_0\Lambda$. Choosing $\mathbf{U}$ such that it diagonalizes the symmetric matrix Λ proves the first statement. To show uniqueness, we estimate the two terms from (2.18) separately. To the first term, the Courant-Fisher theorem [179] applies componentwise to give the estimate $\langle \mathcal{A}\delta\Phi, \delta\Phi\rangle \geq N \cdot \overline{\lambda}\|\delta\Phi\|^2_{L^2}$. For the second, choosing $\mathbf{U} = (u_{i,j})_{i,j=1}^N \in O(N)$ so that $\mathbf{U}^T\Lambda\mathbf{U} = \mathrm{diag}(\lambda_i)_{i=1}^N$, where λ_i are the lowest N eigenvalues of A, gives

$$\begin{aligned}
\langle \delta\Phi\Lambda, \delta\Phi\rangle \quad &= \quad \langle \delta\Phi(\mathbf{U}\mathbf{U}^T\Lambda\mathbf{U}\mathbf{U}^T), \delta\Phi\rangle \quad := \quad \sum_{i=1}^{N}\langle \sum_{j=1}^{N} u_{j,i}\lambda_j\delta\varphi_j, \sum_{k=1}^{N} u_{k,i}\delta\varphi_k\rangle \\
&= \quad \sum_{j,k=1}^{N} \lambda_j\delta_{j,k}\langle \delta\varphi_j, \delta\varphi_k\rangle \quad \leq \quad N \cdot \lambda_N\|\delta\Phi\|^2_{L^2}.
\end{aligned}$$

The assertion now follows from Corollary 1.24, which together with Lemma 2.5 implies that the second order condition (2.19) is fulfilled.

2.3 Convergence analysis for a "direct minimization" algorithm

We will now introduce a steepest descent algorithm for the constrained minimization problem on $\mathcal{G}$. After some comments on its application to electronic structure and eigenvalue problems are given in part (ii), the main results are compiled in part (iii). The proofs of those results are then given in (iv).

(i) Gradient algorithm on $\mathcal{G}$. We will use the steepest descent algorithm introduced in Figure 2.1. It may be viewed as an Euler procedure on the manifold $\mathcal{G}$, applied to the differential equation determining the gradient flow [91] on $\mathcal{G}$, see also [2, 3, 98, 123] for background material and [191] for further comments. Because of the non-differential stepsize, the manifold $\mathcal{G}$ is left in each iteration step. Therefore, a projection on the admitted set is performed afterwards in each step. Note also that for the convergence properties of the algorithm, the role of the preconditioners $\mathcal{B}_n^{-1}$ is crucial, see the following remarks.

ALGORITHM: DIRECT MINIMIZATION/PROJECTED GRADIENT DESCENT

Given *initial iterate $\Phi^{(0)} \in V$, evaluation of $\mathcal{J}'(\Phi^{(n)})$ and of preconditioner(s) B_n^{-1},*
Loop over
(1) *Update* $\Lambda^{(n)} := \langle \mathcal{J}'(\Phi^{(n)}), \Phi^{(n)} \rangle \in \mathbb{R}^{N\times N}$,
(2) *Let* $\hat{\Phi}^{(n+1)} := \Phi^{(n)} - \mathcal{B}_n^{-1}(\mathcal{J}'(\Phi^{(n)}) - \Phi^{(n)}\Lambda^{(n)})$,
$\big(= \Phi^{(n)} - \mathcal{B}_n^{-1}(\mathcal{A}_{\Phi^{(n)}}\Phi^{(n)} - \Phi^{(n)}\Lambda^{(n)})$ *for the case that* $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi\Phi$.$\big)$
(3) *Let* $\Phi^{(n+1)} = P\hat{\Phi}^{(n+1)}$ *by projection P onto $\mathcal{G}$*
until convergence.

Figure 2.1: A "direct minimization" algorithm on $\mathcal{G}$.

Remarks 2.11. (Comments on Algorithm 2.1)

(i) For the applications from Section 2 we have in mind (see the following comments in part (ii) of this section), the gradient $\mathcal{J}'(\Phi)$ is given by $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi\Phi$ with A_Φ the Fock operator, the Kohn-Sham operator or a fixed operator $A_\Phi = A$. Therefore,

$$(\mathcal{J}'(\Phi^{(n)} - \Phi^{(n)}\Lambda^{(n)}))_i \quad = \quad A_{\Phi^{(n)}}\varphi_i^{(n)} - \sum_{j=1}^{N}\langle A_{\Phi^{(n)}}\varphi_i^{(n)}, \varphi_j^{(n)}\rangle\varphi_j^{(n)}$$

is the usual "subspace residual" of the iterate $\Phi^{(n)}$.

(ii) The projection onto $\mathcal{G}$ performed in step (3) only has to satisfy

$$\text{span } \{\varphi_i^{(n+1)} : i \in N\rfloor\} = \text{span } \{\widehat{\varphi}_i^{(n+1)} : i \in N\rfloor\};$$

therefore, any orthogonalization of $\{\widehat{\varphi}_i^{(n+1)} : i \in N\rfloor\}$ is admissible. For example, three favorable possibilities which up to unitary transformations yield the same result are

(a) Gram-Schmidt orthogonalization,

(b) Diagonalization of the Gram matrix $\mathbf{G} = (\langle \hat{\varphi}_i^{(n+1)}, \hat{\varphi}_j^{(n+1)} \rangle)_{i,j=1}^N$ by Cholesky factorization,

(c) For the problems of Section 2.1, i.e. where $\mathcal{J}'(\Phi) = \mathcal{A}_\Phi \Phi$, diagonalization of the matrix

$$\mathbf{A}_{\Phi^{(n+1)}} := (\langle A_{\Phi^{(n)}} \hat{\varphi}_i^{(n+1)}, \hat{\varphi}_j^{(n+1)} \rangle)_{i,j=1}^N$$

by solving an $N \times N$ eigenvalue problem.

(iii) The preconditioner $\mathcal{B}_n^{-1}$ used in the n-th step is induced (via (2.10)) by an elliptic symmetric operator $B_n : V \to V'$, which we require to be equivalent to the norm on V in the sense that [29]

$$\langle B_n \varphi, \varphi \rangle_{L^2(\Omega)} \quad \sim \quad \|\varphi\|^2_{H^1(\Omega)} \quad \forall \varphi \in V. \tag{2.24}$$

(iv) To guarantee convergence of the algorithm, the preconditioners B_n has to be scaled properly by a factor $\alpha > 0$, cf. Lemma 2.17. The optimal choice of α is provided by minimizing the corresponding functional over span $\{\Phi^{(n)}, \widehat{\Phi}^{(n+1)}\}$ (a line search over this space), which can be done for the simplified problem without much additional effort. For the HF-/KS-energy functional, it will become prohibitively expensive. Instead, subspace acceleration techniques like DIIS (see Section 4) provide an attractive alternative to improve the convergence speed. Note that although we will show below that it suffices to fix a suitable parameter α, one might as well use different step sizes for every entry, i.e. $\mathcal{B}_n \Phi = (\alpha_1 B_n \varphi_1, \ldots, \alpha_N B_n \varphi_N)$.

□

(ii) Direct minimization - applications in electronic structure calculations. The above algorithm is the so-called *direct minimization* scheme utilized in HF/DFT calculation, which performs a steepest descent algorithm by updating the gradient of $\mathcal{J}$, i.e. the Kohn-Sham Hamiltonian or Fock operator, in each iteration step. Direct minimization, as proposed in [5], is prominent in DFT calculations if good preconditioners are available and the systems under consideration are large, e.g. for the computation of electronic structure in bulk crystals using plane waves, finite differences [21] and the recent wavelet code developed in the BigDFT project [83].

[29] For DFT/HF calculations, one can use approximations of the shifted Laplacian, $B \approx \alpha(-\frac{1}{2}\Delta + C)$, as is done in the BigDFT project [83]. This is also a suitable choice when dealing with plane wave ansatz functions using advantages of FFT, or a multi-level preconditioner if one has finite differences, finite elements or multi-scale functions like wavelets [9, 21, 86, 96]. For CI, the standardly used preconditioner is the (in canonical orbitals diagonal) Fock operator F, see Section 2.2(iv).
For the simplified problem, the choice $B^{-1} = \alpha A^{-1}$ corresponds to a variant of simultaneous inverse iteration. The choice $B = \alpha(A - \lambda_j^{(n)} I)|_{V_0^\perp}$, where $V_0^\perp := \{v | \langle v, \varphi_i^{(n)} \rangle = 0 \text{ for all } i \in N]\}$, corresponds to a simultaneous Jacobi-Davidson iteration.

In contrast to the direct minimization procedure is the *self consistent field iteration (SCF)*, which keeps the Fock operator fixed until convergence of the corresponding eigenfunctions and updates the Fock operator with the computed eigenbasis thereafter. Note that this means that in the inner iteration loo, the simpler Problem 2.2 is solved for $A = F^{(n)}$; therefore, the above Projected Gradient Descent Algorithm also provides a reasonable routine for the solution of the *inner problem* of SCF, and the results presented here apply to the inner routines of that problem, cf. Lemma 2.9, Remark 2.8. On the whole though, SCF is faced with convergence problems, which have to be remedied by advanced techniques [42] to guarantee convergence. Because the direct minimization scheme with its favourable convergence properties shown below differs from SCF only in that the Fock operator is updated after each inner iteration step, it should be preferred if the update of the Fock operator is sufficiently cheap, which is mostly the case for Gaussians and, by use of *magic filter* techniques [83], for wavelet bases, but not for plane wave bases or finite difference schemes.
For the simpler Problem 2.2, the above algorithm is a multiple-eigenvalue version of the Preconditioned Inverse Iteration scheme that has for the case $N = 1$ extensively been analysed [34, 58, 65, 120, 121, 122, 156, 157, 181, 188], and convergence behaviour is robust in practice. See also [159] for an analysis of the subspace case.

Remark 2.12. (Møller-Plesset perturbation theory)

Let us remark at this point that the non-iterative perturbational ansatz *MP2* [103], which is often applied to improve an energy obtained from a Hartree-Fock solution when a post-HF calculation is too computationally costly, coincides with the first step of the above direct minimization algorithm applied to the CI method, if the Hartree-Fock solution $\Psi_{HF,k}$ is used as starting value and the lifted, shifted Fock operator $F_{HF} - \Lambda_0 I$ (see Remark 3.14) is taken as preconditioner. Thus, for $\Psi_{HF,k}$ sufficiently close to the real solution $\underline{\Psi}$, the below results hold also for the MP2 procedure applied to $\Psi_{HF,k}$ in the sense that MP2 then provides an improved approximation to $\underline{\Psi}$. Note also that higher order variants MPn, $n > 2$, of Møller-Plesset perturbation theory do not allow for such an interpretation, and there are cases known where the MPn energy diverges as a function of n. Also see e.g. [201] for a nice general introduction to how perturbation theory is used in electronic structure theory.

(iii) Convergence analysis: Assumptions and main results. We will now analyse the convergence properties of the above Projected Gradient Descent Algorithm. Recall that in our framework introduced in the beginning of this section, we kept the freedom of choice to either use $V := H^1(\Omega)$, equipped with an inner product equivalent to the H^1-inner product $\langle \cdot, \cdot \rangle_{H^1}$, for analysing the original equations, or to use a finite dimensional subspace $V_d \subset H^1(\Omega)$ for a corresponding Galerkin discretisation of these equations. In practice, our iteration scheme is only applied to the discretised equations. However, the convergence estimates obtained will be uniform with respect to the discretisation parameters. Our analysis bases on the following condition imposed on the functional $\mathcal{J}$.

Assumption 2.13. *Let Φ_0 be a minimizer of (2.1). The second order derivative of the Lagrangian $\mathcal{L}(\Phi_0, \Lambda)$ with respect to Φ_0 is assumed to be V^N-elliptic on the tangent space, i.e. there is $\gamma > 0$ so that*

$$\langle\langle \mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda)\delta\Phi \ , \ \delta\Phi\rangle\rangle \quad \geq \quad \gamma \ \|\delta\Phi\|^2_{V^N}, \qquad \textit{for all} \quad \delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G}. \tag{2.25}$$

From Section 2.2, we recall that $\mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda)\Phi = \mathcal{J}''(\Phi_0)\Phi - \Phi\Lambda$, so that (2.25) is verified if and only if for $\Lambda = (\langle\langle(\mathcal{J}'(\Phi_0))_j, \psi_i\rangle\rangle)_{i,j=1}^N$ as above

$$\langle\langle \mathcal{J}''(\Phi_0)\delta\Phi - \delta\Phi\Lambda \ , \ \delta\Phi\rangle\rangle \quad \geq \quad \gamma \ \|\delta\Phi\|^2_{V^N}, \qquad \text{for all} \quad \delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G} \tag{2.26}$$

holds. Note again that for Hartree-Fock calculations, verification of $\mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda) > 0$ on $\mathcal{T}_{[\Phi_0]}\mathcal{G}$ already implies (2.25), cf [146]. From the present state of Hartree-Fock theory, it is not possible to decide whether this condition is true in general; the same applies for DFT theory. For the simpler eigenvalue problem, the condition holds if the operator $A : V \to V'$ is a bounded symmetric operator, fulfilling the Gårding inequality (1.66) and the gap condition

$$\lambda_N \quad < \quad \inf\{\lambda \mid \lambda \in \sigma(A)\backslash\{\lambda_1, \ldots, \lambda_N\}\}, \tag{2.27}$$

see Lemma 2.9. To formulate our main convergence result, we now introduce a norm $\|.\|_{V^N}$ on the space V^N, which will be equivalent to the $(H^1)^N$-norm but more convenient for our proof of convergence.
Let $B : V \to V'$ be the preconditioning mapping introduced in (i) of this section, so that in particular, B is symmetric and the spectral equivalence

$$\vartheta\|x\|^2_V \quad \leq \quad \langle Bx, x\rangle \quad \leq \quad \Theta\|x\|^2_V$$

holds for some $0 < \vartheta \leq \Theta$ and all $x \in V$. Let us consider the mapping

$$\hat{B}^{-1} : V' \to V \ , \quad \hat{B}^{-1} := (I - D)B^{-1}(I - D) + D, \tag{2.28}$$

where $D = D_{\Phi_0}$ projects onto the sought subspace. Then the inverse $\hat{B}$ satisfies $\langle\hat{B}\varphi, \psi\rangle = \langle\varphi, \hat{B}\psi\rangle$ for all $\varphi, \psi \in V$. Because $\hat{B}^{-1}$ agrees with B^{-1} up to a B^{-1}-compact perturbation [206], there holds for the induced $\hat{B}$-norm $\|.\|_{\hat{B}}$ on V that

$$\langle\hat{B}\varphi, \varphi\rangle \quad \sim \quad \|\varphi\|^2_V.$$

Using the notation (2.10), a norm on V^N is now induced by the $\|.\|_{\hat{B}}$-norm by

$$\|\Phi\|^2_{V^N} := \langle\langle\widehat{\mathcal{B}}\Phi, \Phi\rangle\rangle. \tag{2.29}$$

If we denote by $\Psi(\Phi) \in \mathbb{H}^1_k$ the Slater determinant formed from the N functions contained in Φ, it is not hard to show that $\|\Psi(\Phi) - \Psi(\Phi')\|_{H^1} \lesssim \|\Phi - \Phi'\|_{V^N}$ for any $\Phi, \Phi' \in V^N$, so that estimates for the convergence of $\Phi \in V$ also imply estimates in the original tensor

space $\mathbb{H}^1_k$. The norm (2.29), as any norm defined on V^N in the above fashion, is invariant under the orthogonal group of $\mathbb{R}^{N\times N}$ in the sense that

$$\|\Phi \mathbf{U}\|_{V^N} = \|\Phi\|_{V^N} \tag{2.30}$$

for all $\mathbf{U} \in O(N)$. In the Grassmann manifold, we measure the error between $[\Phi_{(1)}], [\Phi_{(2)}] \in \mathcal{G}$ by a related metric d given by

$$d(\ [\Phi_{(1)}], [\Phi_{(2)}]\ ) \ := \ \inf_{\mathbf{U}\in O(N)} \|\Phi_{(1)} - \Phi_{(2)}\mathbf{U}\|_{V^N}.$$

If $[\Phi_{(2)}]$ is sufficiently close to $[\Phi_{(1)}] \in \mathcal{G}$, it follows from Lemma 2.6 that this measure given by d is equivalent to the expression

$$\|(\mathcal{I} - \mathcal{D}_{\Phi_{(1)}})\Phi_{(2)}\|_{V^N}, \tag{2.31}$$

in which we used the L_2-orthogonal projector $\mathcal{D}_{\Phi_{(1)}}$ onto the subspace spanned by $\Phi_{(1)}$. In the following, let us use the abbreviation $D = D_{\Phi_0}$ for the projector on the sought subspace, whereever no confusion can arise. In terms of the error measure $\|(\mathcal{I}-\mathcal{D})\Phi\|_{V^N}$, our main convergence result is the following.

Theorem 2.14. *(Local linear convergence of the gradient algorithm)*

Under Assumption (2.25) and for $\Phi^{(0)} \in U_\delta(\Phi_0)$ sufficiently close to Φ_0, there is a constant $\chi < 1$ such that for all $n \in \mathbb{N}_0$,

$$\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \quad \leq \quad \chi \cdot \|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} \tag{2.32}$$

holds for the iterates of the direct minimization algorithm introduced in part (i).

For the Rayleigh quotient $R(\varphi^{(n)})$, i.e. for the simplified problem and $N = 1$, it is known that $R(\phi^{(n)}) - R(\psi) \lesssim \|\psi - \phi^{(n)}\|_V^2$. The next result shows that this property (sometimes called "quadratic convergence of the eigenvalues" in a slight abuse of nomenclature) also holds for the computed energies in the more general case, provided that the constraints are satisfied exactly and the functional is sufficiently often differentiable. The latter is true for Hartree-Fock and the simplified problem, since they both depend polynomially on Φ; for DFT, the properties of the exchange correlation potential are not explicitly fixed, so the question remains open in general in this case.

Theorem 2.15. *("Quadratic convergence" of the energies)*

Suppose that (2.25) holds, that $\mathcal{J}$ is two times differentiable on a neighbourhood $U_\delta(\Phi_0) \subseteq V^N$ of the minimizer Φ_0, and that for fixed $\Phi \in U_\delta(\Phi_0)$, $\mathcal{J}''$ is continuous on the connection line $\{t\Phi_0 + (1-t)\Phi | t \in [0,1]\}$. Then,

$$\mathcal{J}(\Phi) - \mathcal{J}(\Phi_0) \quad \lesssim \quad \|(I - \mathcal{D})\Phi\|_{V^N}^2 \ . \tag{2.33}$$

For the proof of the previous theorems, the following result will be useful. We included it into the main results because it also shows that the "residual" $(\mathcal{I}-\mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})$ may be utilized for practical purposes to estimate the norm of the error $(I-D)\Phi^{(n)}$. For more sophisticated goal-oriented error estimators in the context of Hartree-Fock/DFT calculations, see [200].

Lemma 2.16. *(Residual estimators)*

For δ sufficiently small, there are constants $c, C > 0$ such that if $\|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|_{\hat{B}} < \delta$,

$$c\|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|_{V^N} \quad \leq \quad \|(\mathcal{I}-\mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'} \quad \leq \quad C\|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|_{V^N}, \tag{2.34}$$

i.e. the projected gradient is asymptotically an efficient and reliable error estimator for the subspace error. An analogous result holds for gradient error $\|(\mathcal{I}-\mathcal{D})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'}$. In particular,

$$\mathcal{J}(\Phi^{(n)}) - \mathcal{J}(\Phi_0) \quad \lesssim \quad \|(\mathcal{I}-\mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})\|^2_{(V^N)'}. \tag{2.35}$$

Proof of Lemma 2.16. Let us choose $\bar{\Phi}_0 \in [\Phi_0]$ according to Lemma 2.6 (applied to $\Phi = \Phi^{(n)}$). Letting $\Delta\Phi_0 := \Phi^{(n)} - \bar{\Phi}_0$, there holds by linearization with $D = D_{\Phi_0}$ and usage of Lemma 2.6 that

$$\begin{aligned} & (\mathcal{I}-\mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)}) \\ = \; & (\mathcal{I}-\mathcal{D})\mathcal{J}'(\Phi_0) + (\mathcal{I}-\mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\bar{\Phi}_0,\Lambda)\Delta\bar{\Phi}_0 + \mathcal{O}(\|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|^2_{V^N}) \\ = \; & (\mathcal{I}-\mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\bar{\Phi}_0,\Lambda)(\mathcal{I}-\mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|^2_{V^N}). \end{aligned}$$

By Assumption 2.13,

$$\|(\mathcal{I}-\mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|_{(V^N)'} \sim \|(\mathcal{I}-\mathcal{D})\Phi^{(n)}\|_{V^N},$$

from which the assertion follows. The statement for $\|(\mathcal{I}-\mathcal{D})\mathcal{J}'(\Phi^{(n)})\|_{(V^N)'}$ follows from the same reasoning by replacing $\mathcal{L}^{(2,\Phi_0)}(\Phi_0,\Lambda)$ by $\mathcal{J}''(\Phi_0)$ in the above. The estimate (2.35) will follow from Theorem 2.15 together with (2.34) once this is proven, see below.

□

(iv) Proof of Theorems 2.14 and 2.15. To start, let us sketch our proceeding for the proof of Theorem 2.14: The mapping $\Phi^{(n)} \mapsto \Phi^{(n)} - \mathcal{B}^{-1}(\mathcal{I}-\mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)})$ is a perturbation of the mapping $\Phi^{(n)} \mapsto \Phi^{(n)} - \mathcal{B}^{-1}(\mathcal{I}-\mathcal{D}_{\Phi_0})\mathcal{J}'(\Phi^{(n)})$, so we will at first show that the latter mapping, when applied to an iterate $\Phi^{(n)} \in U_\delta(\Phi_0)\mathcal{G}$, indeed reduces its error in the tangent space of Φ; here, the ellipticity assumption enters as main ingredient. The second part consists of showing that the remaining perturbation terms (including those resulting from projection on the manifold) are of higher order and thus asymptotically neglectable. As a first auxiliary lemma, we will now formulate a rather general result

about how ellipticity on subspaces can be used to construct a contraction on these spaces and then specialize this to the tangent space at the solution Φ_0 and Assumption 2.13 in the subsequent corollary.

Lemma 2.17. *Let $W \subset G \subset W'$ be a Gelfand triple, $U \subset W$ a closed subspace of W and $S, T' : W \to W'$ two bounded elliptic operators, symmetric with respect to the G-inner product $\langle ., .. \rangle_G$, satisfying*

$$\gamma \|x\|_W^2 \quad \leq \quad \langle Sx, x \rangle_G \quad \leq \quad \Gamma \|x\|_W^2, \tag{2.36}$$

and

$$\vartheta \|x\|_W^2 \quad \leq \quad \langle T'x, x \rangle_G \quad \leq \quad \Theta \|x\|_W^2 \tag{2.37}$$

for all $x \in U$. Moreover, let S, T' both map the subspace U to itself. Then there exists a scaled variant $T = \alpha T'$, where $\alpha > 0$, and a constant $\beta < 1$ for which

$$\|(I - T^{-1}S)x\|_T \quad \leq \quad \beta \ \|x\|_T \tag{2.38}$$

for all $x \in U$, where $\|x\|_T^2 := \langle Tx, x \rangle_G$ is the inner product induced by T.

Proof. It is easy to verify that for $\beta := (\Gamma\Theta - \gamma\vartheta)/(\Gamma\Theta + \gamma\vartheta) < 1$ and $\alpha := \frac{1}{2}(\Gamma/\vartheta + \gamma/\Theta)$ there holds

$$|\langle (I - T^{-1}S)x, x \rangle_T| \quad \leq \quad \beta \ \|x\|_T^2 \quad \text{ for all } x \in U. \tag{2.39}$$

Due to the symmetry of T, S as mappings $U \to U$, the result (2.38) follows, see e.g. [95]. □

Let $\lambda_i, i = 1, \ldots, N$ be the lowest eigenvalues of A, $\psi_i, i = 1, \ldots, N$, the corresponding eigenfunctions, and

$$V_0 = \text{span } \{\psi_i : i = 1, \ldots, N\}. \tag{2.40}$$

By Lemma 2.5, there holds for $\Phi_0 = (\psi_1, \ldots, \psi_N)$, $V_0 = \text{span}\{\psi_1, \ldots, \psi_N\}$ that $(V_0^\perp)^N = \mathcal{T}_{[\Phi_0]}\mathcal{G}$. The following corollary is the main result needed for estimation of the linear part of the iteration scheme.

Corollary 2.18. *(Contraction property on the tangent space)*

Let $\mathcal{J}$ fulfil the ellipticity condition (2.25) and $B' : V \to V'$ a symmetric operator that fulfils (2.37) with $T' = B'$. Then there exists a scaled variant $B = \alpha B'$, where $\alpha > 0$, for which for any $\delta\Phi \in \mathcal{T}_{[\Phi_0]}\mathcal{G}$ there holds

$$\|\delta\Phi - \hat{\mathcal{B}}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\Phi_0, \Lambda)\delta\Phi\|_{V^N} \quad \leq \quad \beta \ \|\delta\Phi\|_{V^N},$$

where $\beta < 1$ and $\hat{B}$ is defined by B via (2.28).

Proof. Note that the restriction of $\hat{B}'$ is a symmetric operator $V_0^\perp \to V_0^\perp$. Therefore, the extension $\hat{\mathcal{B}}'$ is also symmetric as mapping $\mathcal{T}_{[\Phi_0]}\mathcal{G} \to \mathcal{T}_{[\Phi_0]}\mathcal{G}$. Also, $(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Phi_0)}$ maps $V_0^\perp \to V_0^\perp$ symmetrically, so Lemma 2.17 applies.

□

The last ingredient for our proof of convergence is the following lemma which will imply that the projection following each application of the iteration mapping does not destroy the asymptotic linear convergence.

Lemma 2.19. *(Effects of orthogonalization)*
Let $\hat{\Phi}^{(n+1)} = (\hat{\varphi}_1, \dots, \hat{\varphi}_N)$ be the intermediate iterates as resulting from iteration step (2) in algorithm 1 or 2, respectively. For any orthonormal set $\Phi \in \mathcal{V}$ fulfilling $\mathrm{span}[\Phi] = \mathrm{span}[\hat{\Phi}^{(n+1)}]$*, its error deviates from that of $\hat{\Phi}^{(n+1)}$ only by quadratic error term:*

$$\|(\mathcal{I} - \mathcal{D})\Phi\|_{V^N} \quad = \quad \|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n+1)}\|_{V^N} \; + \; \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n)}\|_{V^N}^2). \tag{2.41}$$

□

Proof. First of all, note that if (2.41) holds for one orthonormal set Φ with $\mathrm{span}[\Phi] = \mathrm{span}[\hat{\Phi}^{(n+1)}]$, it holds for any other orthonormal set $\tilde{\Phi}$ with $\mathrm{span}[\tilde{\Phi}] = \mathrm{span}[\hat{\Phi}^{(n+1)}]$ because

$$\|(\mathcal{I} - \mathcal{D})\Phi\mathbf{U}\|_{V^N} = \|(\mathcal{I} - \mathcal{D})\Phi\|_{V^N}$$

for all orthonormal $\mathbf{U} \in O(N)$. Therefore, we will show (2.41) for $\Phi = (\varphi_1, \dots, \varphi_N)$ yielded from $\hat{\Phi}^{(n+1)}$ by the Gram-Schmidt orthonormalization procedure. Denote $\hat{\varphi}_i = \varphi_i^{(n)} + r_i^{(n)}$, where

$$r_i^{(n)} \quad = \quad (B^{-1}\big(\mathcal{I} - \mathcal{D}_{\Phi^{(n)}})\mathcal{J}'(\Phi^{(n)}))_i.$$

From the previous lemma, we get in particular that

$$\|r_i^{(n)}\|_V \lesssim \|(I - D)\varphi_i^{(n)}\|_V$$

(remember that $D = D_{\Phi_0}$). With the Gram-Schmidt procedure given by

$$\varphi_k' \quad = \quad \hat{\varphi}_k \; - \sum_{j<i} \langle \hat{\varphi}_k, \varphi_j \rangle \; \varphi_j, \qquad \varphi_k = \varphi_k'/\|\varphi_k'\|,$$

the lemma is now proven by verifying that in each of the inner products involved, there occurs at least one residual $\|r_i^{(n)}\|$; and that, on top of this, for the correction directions φ_j there holds

$$(I - D)\varphi_j' \quad = \quad \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}) + \mathcal{O}(\sum_{i<k} \|r_i^{(n)}\|_{V^N}) \quad = \quad \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}).$$

Therefore, the correction terms are of $\mathcal{O}(\|(\mathcal{I} - \mathcal{D})\hat{\Phi}^{(n)}\|_{V^N}^2)$, thus proving $\varphi_k' - \hat{\varphi}_k = \mathcal{O}(\|(I - D)\Phi\|_{V^N}^2)$. It is easy to verify that the normalization of φ_k' only adds another quadratic term, so the result follows.

□

To finally prove (2.32), we define $\mathcal{F}(\Phi) = \Phi - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D}_\Phi)\mathcal{J}'(\Phi)$, so that $\Phi^{(n+1)} = P(\mathcal{F}(\Phi^{(n)}))$, where P is a projection on $\mathcal{G}$ for which $\mathrm{span}P(\mathcal{F}(\Phi^{(n)})) = \mathrm{span}\mathcal{F}(\Phi^{(n)})$. For fixed n, let us choose $\bar{\Phi}_0 \in \mathrm{span}[\Phi_0]$ according to Lemma 2.6, so that, using the abbreviation $\mathcal{D} := \mathcal{D}_{\Phi_0}$,

$$\begin{aligned}\bar{\Phi}_0 - \Phi^{(n)} &= (\mathcal{I} - \mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{L_2^N}) \\ &\leq (\mathcal{I} - \mathcal{D})\Phi^{(n)} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N}).\end{aligned}$$

Introducing $\Delta\Phi_0 := \Phi^{(n)} - \bar{\Phi}_0$, there follows by Lemma 2.19 and linearization

$$\begin{aligned}&\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \\ =\ &\|(\mathcal{I} - \mathcal{D})\mathcal{F}(\Phi^{(n)})\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N}) \\ =\ &\|(\mathcal{I} - \mathcal{D})\mathcal{F}(\bar{\Phi}_0) + (\mathcal{I} - \mathcal{D})\mathcal{F}'(\bar{\Phi}_0)\Delta\Phi_0\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N}) \\ =\ &\|(\mathcal{I} - \mathcal{D})\mathcal{F}'(\bar{\Phi}_0)(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N}) \\ =\ &\|(\mathcal{I} - \mathcal{D})\big(\mathcal{I} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\bar{\Phi}_0, \Lambda)\big)(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} \\ &+\mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N})\end{aligned}$$

where we have used (2.42) and the fact that $(\mathcal{I} - \mathcal{D})\mathcal{F}(\bar{\Phi}_0)$ is zero. The proof is now finished by noticing that

$$\begin{aligned}&(\mathcal{I} - \mathcal{D})\Big(\mathcal{I} - \mathcal{B}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\bar{\Phi}_0, \Lambda)\Big)(\mathcal{I} - \mathcal{D})\Phi^{(n)} \\ =\ &\Big(\mathcal{I} - \hat{\mathcal{B}}^{-1}(\mathcal{I} - \mathcal{D})\mathcal{L}^{(2,\Phi_0)}(\bar{\Phi}_0, \Lambda)\Big)(\mathcal{I} - \mathcal{D})\Phi^{(n)},\end{aligned}$$

so that Corollary 2.18 applies to give

$$\|(\mathcal{I} - \mathcal{D})\Phi^{(n+1)}\|_{V^N} \ \leq\ \vartheta\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N} + \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|^2_{V^N}) \ \leq\ \chi\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N},$$

where $\chi < 1$ for $\|(\mathcal{I} - \mathcal{D})\Phi^{(n)}\|_{V^N}$ small enough to neglect the quadratic term.

□

Proof of Theorem 2.15. Let us choose a representant of the solution Φ^* according to Lemma 2.6. Abbreviating $e = \Phi - \Phi^*$, we can use $\mathcal{J}'(\Phi^*)((\mathcal{I} - \mathcal{D})\Phi) = 0$ to find that

$$\mathcal{J}'(\Phi^*)(e) \ =\ \mathcal{J}'(\Phi^*)((\mathcal{I} - \mathcal{D})\Phi) \ +\ \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2) \ =\ \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2)$$

so that

$$\begin{aligned}\mathcal{J}(\Phi) - \mathcal{J}(\Phi^*) &= \int_0^1 \mathcal{J}'(\Phi^* + se)(e)ds \ +\ \frac{1}{2}\mathcal{J}'(\Phi)(e) \\ &\quad - \frac{1}{2}(\mathcal{J}'(\Phi^*)(e) \ +\ \mathcal{J}'(\Phi)(e)) \ +\ \mathcal{O}(\|(\mathcal{I} - \mathcal{D})\Phi\|^2).\end{aligned}$$

By integration by parts,

$$\frac{1}{2}(f(0) + f(1)) \ =\ \int_0^1 f(t)dt \ +\ \int_0^1 (s - \frac{1}{2})f'(s)ds,$$

we get

$$\mathcal{J}(\Phi)-\mathcal{J}(\Phi^*) \;=\; \frac{1}{2}\langle\langle \mathcal{J}'(\Phi), \Phi-\Phi^*\rangle\rangle \;-\; \int_0^1 (s-\frac{1}{2})\mathcal{J}''(\Phi+se)(e,e)ds \;+\; \mathcal{O}(\|(\mathcal{I}-\mathcal{D})\Phi\|^2).$$

For estimation of the first term on the right hand side, recall from (2.34) that

$$\|(\mathcal{I}-\mathcal{D})\mathcal{J}'(\Phi)\|_{V^N} \;\lesssim\; \|(I-\mathcal{D})\Phi\|_{V^N},$$

and therefore

$$\begin{aligned}\frac{1}{2}\langle\langle \mathcal{J}'(\Phi), \Phi-\Phi^*\rangle\rangle &= \frac{1}{2}\langle\langle (\mathcal{I}-\mathcal{D})\mathcal{J}'(\Phi), (\mathcal{I}-\mathcal{D})\Phi\rangle\rangle \;+\; \mathcal{O}(\|(I-\mathcal{D})\Phi\|^2)\\ &= \mathcal{O}(\|(I-\mathcal{D})\Phi\|^2),\end{aligned}$$

while for the second term, $|\int_0^1 (s-\frac{1}{2})\mathcal{J}''(\Phi+se)(e,e)ds| \;=\; \mathcal{O}(\|e\|^2) = \mathcal{O}(\|(I-\mathcal{D})\Phi\|^2)$ follows from the continuity of $\mathcal{J}''$ and, again, the usage of Lemma 2.6.

□

2.4 Concluding remarks

We have shown that under the ellipticity condition (2.25), the direct minimization algorithm displayed in Fig. 2.1 is locally linearly convergent. Let us note that a verification of the ellipticity condition will also in the context of Hartree-Fock/DFT theory answer other important open problems as for instance, uniqueness of solutions.
In accordance with our theoretical results, the convergence behaviour of the gradient algorithm seems to be quite robust, see e.g. [83] and references given there for numerical examples in the context of quantum chemistry, as well as the numerical examples in [158] for the single eigenvalue case.
We already noted in [191] that it would be desirable to extend the present direct minimization approach to the use of sparse, localized and non-orthogonal orbitals in order to achieve rigorous convergence estimates for linear scaling DFT methods, and this approach is momentarily pursued further in our group [127, 128].

□

3 The continuous Coupled Cluster method

The Coupled Cluster (CC) method was derived around 1960 in the context of atomic physics [50, 51, 131, 203], and later introduced in the context of quantum chemistry (see [49]). It is nowadays the probably most widely applied tool in the calculation of molecular properties when high accuracy is demanded. This is due to various favourable properties of the Coupled Cluster method: The CC method enjoys a wide range of applicability in a black-box style and converges quickly and systematically to the full-CI (Galerkin) energy limit[30] E^{disc} when applied to relatively well-behaved systems as typically C-H-chains, rings, alcohols, cetones and aminoacids are. It usually outperforms the CI method of corresponding scaling, see e.g. [55, 134].[31] The CCSD(T) method [176], which can be applied to small to medium-sized molecules with reasonable computational effort, often provides results which are within the error bars of corresponding practical experiments [137], especially if used in connection with extrapolation schemes where approximation of the ground state energy E^* is calculated in various hierarchical basis sets (e.g. VDZ/VTZ/VQZ) and then extrapolated. In contrast to truncated CI methods, truncated CC has the essential property of being size-consistent [18, 19, 160], making CC the superior tool when describing reaction mechanisms. However, there are situations where the Coupled Cluster method may converge slowly or not at all, if the reference determinant Ψ_0 – usually constructed from a preceding Hartree-Fock calculation – is not sufficiently good, see [103]; multi-reference CC methods (e.g. [29, 164, 168]), remedying this shortcoming, are still in their development. For a review on Coupled Cluster theory, the reader is referred to [17, 132] and the abundance of references given therein, as well as to the article [31] for a broader scope on the applications in physics; for some recent developments, see [26, 46, 135, 154] as well as the references given in Section 3.1.

In spite of the CC method's practical utility and popularity, theoretical results from the mathematical point of view are rather scarce. Only recently a first approach has been undertaken in [190], see also [134]. To outline the results from [190] and the results to be proven below, we remind the reader of the two discretisation steps normally taken when discretising the weak eigenvalue problem Problem 1.12 (also cf. Sec. 1.7): First, a finite subset B^{disc} of a complete one-particle basis $B \subseteq H^1(\mathbb{R}^3)$ is chosen, from which an according tensor basis $\mathbb{B}_k^{\mathrm{disc}} \subseteq \mathbb{B}_k$ of a discretisation $\mathbb{H}_k^{1,\mathrm{disc}} \subseteq \mathbb{H}_k^1$ is constructed as outlined in Section 1.4; the resulting (finite-, yet high-dimensional) space $\mathbb{H}_k^{1,\mathrm{disc}}$ is the "full CI-space" (as in contrast to the "complete CI-space" $\mathbb{H}_k^1$). Afterwards, the set of tensor basis functions $\mathbb{B}_k^{\mathrm{disc}}$ is reduced by certain selection criteria normally associated with the so-called "excitation level" of the basis functions, see Definition 3.2, and the Galerkin projection of the "full CI" CI-/CC-equations onto the selected subspace leads

[30] Confer the remarks at the end of Section 1.

[31] For instance, the CCSD method [18], with its complexity of $\mathcal{O}(N^6)$ the same than that of the related CISD method, can for larger molecules be interpreted as an approximation of the more precise CISDTQ method, which itself scales as $\mathcal{O}(N^{10})$.

to the (canonical, projected) Coupled Cluster method, e.g. termed CISD resp. CCSD if only basis functions corresponding to single and double excitations are included. The analysis in [190] now examines the approximation properties of the projected Coupled Cluster method to the "full CI" solution, and thus provides an analysis of the second discretisation step. On the other hand this first approach, taken mainly to circumvent the problems associated with the formulation of the Coupled Cluster method for the original, infinite-dimensional problem, does not allow for estimates with respect to the true solution $\underline{\Psi} \in \mathbb{H}^1_k$, and thus *a priori* excludes approaches where the size of the underlying one-particle basis B^{disc} is varied. The latter are of interest in the context of error estimation though, especially in view of the fact that convergence of different CC models towards the limit within the full CI-space usually is rather fast, while the convergence of the full-CI solutions $\underline{\Psi}^{disc} \in \mathbb{H}^{1,\text{disc}}_k$ to the continuous limit $\underline{\Psi} \in \mathbb{H}^1_k$ is often rather slow with respect to the size of the underlying one-particle basis set.
In this part of this work, we will therefore formulate the Coupled Cluster equations in a coefficient space reflecting the continuous ("complete CI") space $H^1 := \mathbb{H}^1_k$, and the resulting method will be termed "the continuous Coupled Cluster method". First of all, the continuity properties of cluster operators in the respective function spaces H^1, H^{-1} have to be established (Theorem 3.6), and indeed, this poses the main obstacle in the analysis of the continuous CC method. Once this is done, we will formulate the continuous CC equations and define the continuous CC function f. We prove that f possesses the property of being locally strongly monotone in a neighborhood of the solution t^* (Theorem 3.18); then, techniques from operator theory partly already used in [190] apply to obtain existence/uniqueness and convergence results (Theorem 3.21), and we will prove a goal oriented error estimator [22] for convergence of the energy E^* (Theorem 3.24). Finally, we will indicate how the CC equations can be simplified to obtain computable expressions (Section 3.5) and show convergence of a quasi-Newton method (also formulated in the continuous space) when applied to the CC function.

3.1 Notations, basic assumptions and definitions

For our analysis of the CC method, we fix a (not necessarily orthogonal) spin orbital basis

$$B^\Sigma = \{\chi_P \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \mid P \in \mathcal{I}\} \tag{3.1}$$

of $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$. Motivated by the fact that the Coupled Cluster method is a *Post-Hartree-Fock method* (see Section 2(ii)), we will suppose that from a preliminary calculation, e.g. a HF calculation as outlined in Section 2(i), we have a *reference determinant*

$$\Psi_0 \quad = \quad \Psi^{HF} \quad = \quad \mathcal{Q}\big(\chi_{\overline{1}} \otimes \ldots \otimes \chi_{\overline{k}} \otimes \chi_{\underline{1}} \otimes \ldots \otimes \chi_{\underline{N-k}}\big). \tag{3.2}$$

at hand, approximating the sought eigenfunction $\underline{\Psi}$ to a certain extent, and that this rank-1-approximation can be formed from N functions from B^Σ. Ψ_0 is a Slater determinant

of spin $\zeta_k = -N/2 + k$, which will be fixed in this section. A spin orbital χ_I contained Ψ_0 in will be called *occupied orbital*, and this situation will be abbreviated by $I \in \text{occ}$. Iff $A \notin \text{occ}$, a spin orbital χ_A is called *virtual orbital*, denoted by $A \in \text{virt}$. It is a notational convention that in summations etc., occupied orbitals are labeled by letters $I, J, K, \ldots \in \text{occ}$, virtual orbitals by letters $A, B, C, \ldots \in \text{virt}$, and unspecified orbitals by letters $P, Q, R, \ldots \in \mathcal{I}$, and we will also use this convention here.
For the discrete ("projected") Coupled Cluster method in its simplest form, a (discrete) basis B^Σ of so-called canonical orbitals is in practice provided by diagonalization of the final (discrete) Fock operator $F_{HF} = F_{HF,\Psi^{HF}}$, so that B^Σ is an eigenbasis of the Fock operator, and this discrete setting was analysed in [190]. In the infinite dimensional setting, $F_{\Psi_{HF}}$ does not allow for a complete eigensystem anymore, so that the formulation of the Coupled Cluster method and also the analysis from [190] do not extend straightforwardly to the continuous setting. Also many of the more sophisticated CC schemes are not based on canonical orbitals (i.e. the eigenfunctions of the Fock operator), but use certain localization criteria to rotate the occupied orbitals (to e.g. Foster-Boys-type orbitals [33], Pipek-Mazay-type orbitals [172] or enveloped localized orbitals [13]), use non-orthogonal bases for the virtual orbitals (e.g. the projected atomic orbitals (PAOs) in the LCCSD approach [100, 194]), or enhance the virtual space by specialized basis functions taking the electron-electron cusp in account (as e.g. the recent powerful $r_{1,2}$- and $f_{1,2}$- methods [119]). Nevertheless, if a HF ground state exists (see 2.1(i)), the infinite dimensional Fock operator F_{HF} possesses an invariant subspace belonging to N lowest eigenvalues, and the L_2- and F_{HF}-orthogonality between virtual and occupied orbitals is maintained in all of the aforementioned methods. Motivated by this, we will base our analysis on the following mild assumptions covering all of the above cases.

Assumption 3.1. *We have a symmetric mapping*

$$F : H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}) \to H^{-1}(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$$

at hand that induces a norm spectrally equivalent to the $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$-norm, i.e. there are $\gamma, \Gamma > 0$ such that

$$\gamma \, \langle \varphi, \varphi \rangle_1 \quad \leq \quad \langle F\varphi, \varphi \rangle \quad \leq \quad \Gamma \, \langle \varphi, \varphi \rangle_1 \qquad \text{for all } \; \varphi \in H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\}). \tag{3.3}$$

For the spin basis (3.1), we will suppose that χ_P are eigenfunctions of the z-spin operator S_N^z, see Section 1.2(ii). We also demand that $\{\chi_I | I \in \text{occ}\}$ is a basis of an invariant subspace of F, that is, there holds

$$\langle F\chi_I, \chi_A \rangle \; = \; \langle \chi_I, \chi_A \rangle \; = \; 0 \quad \text{for all } \; I \in \text{occ}, \; A \in \text{virt}\,. \tag{3.4}$$

□

By Lemma 1.25, the above mapping F induces a norm on the tensor space $\mathcal{L}^2_{\mathbb{R}}$ that is equivalent to the $\mathcal{H}^1$-norm. Note that the condition (3.3) is in particular fulfilled by the (continuous or discrete) shifted Fock operator $F = F_{HF} - \mu I$ and also by suitable Kohn-Sham Hamiltonians, see Lemma 1.23 and Remark 2.8.
We will in this section abbreviate by $H^1 := \mathbb{H}^1_k$, $L^2 := \mathbb{L}^2_k$ the corresponding spaces of real-valued, antisymmetric functions of fixed spin number $0 \leq k \leq N/2$ introduced in Section 1.2, and rewrite the weak Schrödinger equation with the dual argument to the right,

$$\langle \Psi, (\hat{H} - E^*)\underline{\Psi} \rangle \;=\; 0 \quad \text{for all} \quad \Psi \in H^1 := \mathbb{H}^1_k, \tag{3.5}$$

to stay consistent with other literature on the Coupled Cluster method. For convenience, we will impose to the solution of the weak Schrödinger equation the *intermediate normalization condition*, i.e. we drop the normalization condition (1.11) and instead we look for eigenfunctions $\underline{\Psi} = \Psi_0 + \Psi^*$ for which

$$\langle \Psi_0, \Psi^* \rangle = 0, \quad \text{i.e.} \quad \langle \underline{\Psi}, \Psi_0 \rangle = 1 \tag{3.6}$$

is fulfilled. This poses no additional restriction if the reference solution is sufficiently good so that $\langle \underline{\Psi}, \Psi_0 \rangle \neq 0$, and we assume this latter condition from now on. The eigenfunction $\underline{\Psi}$ is thus fixed by its component $\Psi^* \in \operatorname{span}\{\Psi_0\}^\perp$, and we will now, as a first step towards the CC formulation, rewrite Ψ^* in terms of so-called excitations of the reference determinant Ψ_0. We start with some definitions.

Definition 3.2. *(Operator strings and excitation operators)*
Let $b_1, \ldots, b_n$ be any canonical creation or annihilation operators. An operator of the form $S : \mathbb{F} \to \mathbb{F}, S\Psi = b_1 \ldots b_n \Psi$, will be called *operator string*.
An operator string

$$X^{A_1,\ldots,A_r}_{I_1,\ldots,I_r} = a^\dagger_{A_1} \ldots a^\dagger_{A_r} a_{I_1} \ldots a_{I_r} \tag{3.7}$$

is called *excitation operator* if $I_1 < \ldots < I_r \in \text{occ}$, $A_1 < \ldots < A_r \in \text{virt}$, and if in $\{I_1, \ldots, I_r\}$ and $\{A_1, \ldots, A_r\}$, the numbers of contained "spin up" indices coincide. The number $r = r(X^{A_1,\ldots,A_r}_{I_1,\ldots,I_r}) \leq N$ of annihilators (resp. creators) contained in $X^{A_1,\ldots,A_r}_{I_1,\ldots,I_r}$ is called the *(excitation) rank* of $X^{A_1,\ldots,A_r}_{I_1,\ldots,I_r}$.

□

Definition 3.3. *(Indices and index operations)*

(i) We denote by $\mu_0 := (\overline{1}, \ldots, \overline{k}, \underline{1}, \ldots, \underline{N-k})$ the index belonging to the reference determinant, and let

$$\mathcal{M}^* = \mathcal{M} \backslash \{\mu_0\}, \qquad \mathcal{M}^*_k = \mathcal{M}_k \backslash \{\mu_0\}$$

with the index sets $\mathcal{M}, \mathcal{M}_k$ from Definition 1.18 .

(ii) For a multi-index $\mu \in \mathcal{M}_k^*$, corresponding to an excitation operator $X_\mu = X_{I_1,\dots,I_r}^{A_1,\dots,A_r} = a_{A_1}^\dagger \dots a_{A_r}^\dagger a_{I_1} \dots a_{I_r}$ and a determinant $\Psi_\mu = \Psi_{I_1,\dots,I_r}^{A_1,\dots,A_r}$, we define its rank as $r(\mu) := r(X_\mu)$.
Iff $P \in \{I_1, \dots, I_r, A_1, \dots, A_r\}$ we say that P is contained in μ, $P \in \mu$ in short. For μ_0, we define that $P \notin \mu_0$ for all $P \in \mathcal{I}$.

(iii) For two multi-indices $\nu, \mu \in \mathcal{M}$, we write $\mu \subseteq \nu$ iff for all indices $P \in \mathcal{I}$, $P \in \mu$ implies $P \in \nu$.

(iv) Obviously, for each pair $\mu \subseteq \nu \in \mathcal{M}_k$, there is exactly one multi-index $\alpha \subseteq \nu \in \mathcal{M}_k$ determined by the condition that $P \in \alpha$ iff $P \in \nu, P \notin \mu$, and we will denote the relation between these indices by $\nu = \mu \oplus \alpha$, $\alpha = \nu \ominus \mu$.
Additionally, we define for the situations where $\oplus, \ominus$ is not defined by the above that $\mu \oplus \alpha = -1$ if $\{P|P \in \mu\} \cap \{P|P \in \alpha\} \neq \emptyset$, and $\nu \ominus \mu = -1$ for the case $\mu \not\subseteq \nu$.

(v) Finally, we declare for convenience that $X_{\mu_0} = I$, define that for coefficients turning up in summations etc. $c_{-1}, t_{-1}, \dots = 0$, and also let $\Psi_{-1} = 0, X_{-1} = 0$.

□

Remarks 3.4. (Properties of determinants and excitation operators)

(i) An excitation operator $X_{I_1,\dots,I_r}^{A_1,\dots,A_r}$ maps the reference determinant $\Psi_0 \in \mathbb{B}_k$ (of fixed spin number k) by definition to a Slater determinant $\Psi_\mu \in \mathbb{B}_k$ by replacing the occupied orbitals $I_1, \dots, I_r$ by the virtual orbitals $A_1, \dots, A_r$. More precisely, we have a one-to-one correspondence between the basis functions $\Psi_\mu \in \mathbb{B}_k$ and the excitation operators $X_{I_1,\dots,I_r}^{A_1,\dots,A_r}$, and because both notations will be convenient in some situations, we will identify the index sets and therefore write $\Psi_\mu = \Psi_{I_1,\dots,I_r}^{A_1,\dots,A_r} := X_{I_1,\dots,I_r}^{A_1,\dots,A_r} \Psi_0$. Also, we will denote the excitation operator taking Ψ_0 to Ψ_μ by X_μ; further, we will call $\Psi_\mu = \Psi_{I_1,\dots,I_r}^{A_1,\dots,A_r}$ an $r-$fold excited determinant or determinant of excitation rank r. Note also that by Lemma 1.20,$(X_{I_1,\dots,I_r}^{A_1,\dots,A_r})^\dagger = a_{I_1}^\dagger \dots a_{I_r}^\dagger a_{A_1} \dots a_{A_r}$, so that

$$(X_{I_1,\dots,I_r}^{A_1,\dots,A_r})^\dagger \; X_{I_1,\dots,I_r}^{A_1,\dots,A_r} \; \Psi_0 \quad = \quad (X_{I_1,\dots,I_r}^{A_1,\dots,A_r})^\dagger \; \Psi_{I_1,\dots,I_r}^{A_1,\dots,A_r} \quad = \quad \Psi_0, \tag{3.8}$$

and the adjoints of excitation operators are therefore sometimes termed decitation operators.[32]

(ii) For two determinants Ψ_r, Ψ_s of excitation ranks $r \neq s$,

$$\langle \Psi_r, \Psi_s \rangle = \langle \Psi_r, \Psi_s \rangle_F = 0 \tag{3.9}$$

due to (3.4).

[32]Note that $(X_{I_1,\dots,I_r}^{A_1,\dots,A_r})^\dagger$ is not the inverse of $X_{I_1,\dots,I_r}^{A_1,\dots,A_r}$ though.

(iii) It follows from the anticommutator relations 1.20(v) that all operators contained in any excitation operators anticommute. Therefore, Definition 3.3 implies that for all indices $\alpha, \beta \in \mathcal{M}_k$, $X_{\alpha\oplus\beta}$ also defines an excitation operator, and that

$$X_\alpha X_\beta \;=\; X_{\alpha\oplus\beta} \;=\; X_{\beta\oplus\alpha} \;=\; X_\beta X_\alpha. \tag{3.10}$$

The same holds for products of decitation operators $X_\alpha^\dagger X_\beta^\dagger = X_{\alpha\oplus\beta}^\dagger$. Also,

$$X_\alpha^\dagger X_\beta = X_{\beta\ominus\alpha}, \qquad X_\alpha \Psi_\beta = \Psi_{\beta\oplus\alpha}, \qquad X_\alpha^\dagger \Psi_\beta = \Psi_{\beta\ominus\alpha}. \tag{3.11}$$

□

Observation/Definition 3.5. *(Cluster operator)*
Due to Remark 3.4(i), every intermediately normed $\Psi \in L^2$ can be expanded in the tensor basis $\mathbb{B}_k$ as

$$\Psi \quad = \quad \Psi_0 + \Psi^* \;=\; \Psi_0 + \sum_{\mu\in\mathcal{M}_k^*} t_\mu X_\mu \Psi_0 \;=:\; (I + T_{\Psi^*})\Psi_0 \tag{3.12}$$

of at most N-fold spin-k-excitations $X_\mu\Psi_0$ of the reference determinant $\Psi_0 \in \mathbb{B}_k$. The operator T_{Ψ^*} introduced in the above remark will be called *cluster operator* of $\Psi \in L^2$.

□

3.2 Continuity properties of cluster operators; the Coupled Cluster equations

(i) Continuity properties of cluster operators. This part of this section is devoted to the proof of the following theorem, which is fundamental for the continuous formulation of the Coupled Cluster equations. After it is proven, we establish in part (ii) the exponential parameterisation of the eigenvalue problem (3.5) which then gives rise to the continuous Coupled Cluster equations.

Theorem 3.6. *(L^2-/H^1-continuity of the cluster operator and its adjoint)*
For any $\Psi^* = \sum_{\alpha\in\mathcal{M}^*} t_\alpha \Psi_\alpha$, the cluster operator $T = T_{\Psi^*}$ and its L^2-adjoint $T^\dagger = T_{\Psi^*}^\dagger$ map $L^2 \to L^2$ boundedly; there holds

$$\|T\|_{L^2\to L^2} = \|T^\dagger\|_{L^2\to L^2} \sim \|\Psi^*\|_{L^2}. \tag{3.13}$$

If $\Psi^* \in H^1$, T and $T^\dagger$ also map $H^1 \to H^1$ boundedly, and

$$\|T\|_{H^1\to H^1} \sim \|\Psi^*\|_{H^1}, \qquad \|T^\dagger\|_{H^1\to H^1} \le \|\Psi^*\|_{H^1}. \tag{3.14}$$

In contrast to the proof for (3.13), which is essentially identical to that for the discrete ("projected") setting analysed in [190], the H^1-continuity (3.14) of T and $T^\dagger$ is considerably harder to prove. This is rooted in the fact that we cannot suppose anymore that the preconditioner $F : H^1(\mathbb{R}^3) \to H^{-1}(\mathbb{R}^3)$ fulfilling (3.4) admits a complete eigensystem. We note also that in [190], it was used that the discrete Hamiltonian $\mathbf{H}$ boundedly maps to ℓ_2 for each discretisation, so for definition of the *discrete* Coupled Cluster equations, the need to show the continuity of $T^\dagger : H^1 \to H^1$ could be avoided. This is not the case any more in the continuous setting.[33]
We start the proof of Theorem 3.6 by showing that we can without loss of generality suppose that the spin basis B^Σ, determining Ψ^* and T, is L_2-orthonormal.

Lemma 3.7. *(Reduction to orthonormal basis sets)*

Let $\tilde{B}^\Sigma := \{\tilde{\chi}_I \mid I \in \mathrm{occ}\} \cup \{\tilde{\chi}_A \mid A \in \mathrm{virt}\}$ *be an* L_2*-orthonormal basis for which there holds*

$$\mathrm{span}\{\tilde{\chi}_I | I \in \mathrm{occ}\} = \mathrm{span}\{\chi_I | I \in \mathrm{occ}\}, \quad \mathrm{span}\{\tilde{\chi}_A | A \in \mathrm{virt}\} = \mathrm{span}\{\chi_A | A \in \mathrm{virt}\},$$

and denote by $\tilde{\Psi}_\alpha$ *the elements of the tensor basis constructed from* $\tilde{B}^\Sigma$*, and by* $\tilde{X}_\alpha, \alpha \in \mathcal{M}^*_k$*, the excitation operators constructed from the creators and annihilators belonging to the basis functions from* $\tilde{B}^\Sigma$*.*

(i) There holds $\mathrm{span}\{\Psi_\alpha | \alpha \in \mathcal{M}^*_k\} = \mathrm{span}\{\tilde{\Psi}_\alpha | \alpha \in \mathcal{M}^*_k\}$*.*

(ii) For the cluster operator $T = \sum_{\alpha \in \mathcal{M}^*_k} t_\alpha X_\alpha$ *belonging to*

$$\Psi^* \;=\; \sum_{\alpha \in \mathcal{M}^*_k} t_\alpha \Psi_\alpha \;=\; \sum_{\alpha \in \mathcal{M}^*_k} \tilde{t}_\alpha \tilde{\Psi}_\alpha \;\in\; \mathrm{span}\{\Psi_\alpha | \alpha \in \mathcal{M}^*_k\},$$

there also holds $T = \sum_{\alpha \in \mathcal{M}^*_k} \tilde{t}_\alpha \tilde{X}_\alpha$*.*

Proof. First of all, (3.9) gives that $\langle \Psi_0, \Psi_\alpha \rangle = 0$ and (3.4) implies that $\langle \tilde{\Psi}_0, \Psi_\alpha \rangle = 0$ for all $\alpha \in \mathcal{M}^*_k$, implying $\mathrm{span}\{\Phi_0\} = \mathrm{span}\{\tilde{\Phi}_0\}$ and thus, with (3.9), $\mathrm{span}\{\Psi_\alpha | \alpha \in \mathcal{M}^*_k\} = \mathrm{span}\{\tilde{\Psi}_\alpha | \alpha \in \mathcal{M}^*_k\}$. Let us denote by $\tilde{a}_P, \tilde{a}^\dagger_P$ the annihilator/creator of $\tilde{\chi}_P$, respectively. Again using (3.4), we can expand

$$\chi_I = \sum_{J \in \mathrm{occ}} c_I^J \tilde{\chi}_J, \quad \chi_A = \sum_{B \in \mathrm{virt}} c_A^B \tilde{\chi}_B, \quad a_I = \sum_{J \in \mathrm{occ}} c_I^J \tilde{a}_J, \quad a_A^\dagger = \sum_{B \in \mathrm{virt}} c_A^B \tilde{a}_B^\dagger,$$

where we inserted the expansions for χ_I, χ_A into the representations (1.54) and (1.56) for the creation and annihilation operators. Thus, for suitable coefficients $d_\alpha^{\alpha'}$, $\alpha, \alpha' \in \mathcal{M}^*$,

$$T = \sum_{\alpha \in \mathcal{M}^*} t_\alpha X_\alpha = \sum_{\alpha \in \mathcal{M}^*} t_\alpha \Big(\sum_{\substack{\alpha' \in \mathcal{M}^* \\ \mathrm{rk}(\alpha') = \mathrm{rk}(\alpha)}} d_\alpha^{\alpha'} \Big) \tilde{X}_\alpha = \sum_{\alpha' \in \mathcal{M}^*} \Big(\sum_{\substack{\alpha \in \mathcal{M}^* \\ \mathrm{rk}(\alpha') = \mathrm{rk}(\alpha)}} t_\alpha d_\alpha^{\alpha'} \Big) \tilde{X}_{\alpha'}. \tag{3.15}$$

[33] Note also that the continuity of $T : H^1 \to H^1$ only implies the continuity of its H^1-adjoint $T^{\dagger, H^1} : H^{-1} \to H^{-1}$, but not the H^1-continuity of the L_2-adjoint $T^\dagger : H^1 \to H^1$ of T.

Because

$$\sum_{\alpha\in\mathcal{M}_k^*} \tilde{t}_\alpha \tilde{X}_\alpha \Psi_0 = \Psi^* = T\Psi_0 = \sum_{\alpha'\in\mathcal{M}^*} \Big(\sum_{\substack{\alpha\in\mathcal{M}^* \\ \mathrm{rk}(\alpha')=\mathrm{rk}(\alpha)}} t_\alpha d_\alpha^{\alpha'} \Big) \tilde{X}_{\alpha'} \Psi_0,$$

the coefficients to the very left and the very right coincide, so (ii) follows from (3.15). □

We will now of course use Lemma 3.7 and assume that B^Σ is orthonormal. Using that for each z-spin-eigenvalue $\zeta \in \mathrm{spin}(N)$, the image $T\Psi_\zeta$ of a z-spin-eigenfunction Ψ_ζ belonging to ζ is again an eigenfunctions belonging to ζ, we will prove the continuity of $T, T^\dagger$ as mappings $\mathbb{L}^2 \to \mathbb{L}^2$ (instead of $L^2 \to L^2$) to avoid subtleties in the choice of basis sets. For the proof, we expand $T\Psi$ in suitable orthonormal bases and then estimate the occurring terms by the below Lemma 3.9. We start by introducing some short-hand notations for occurring terms.

Notations 3.8. *(Notations used in the proof of Theorem 3.6)*

(i) The index $\mu \in \mathcal{M}^*$ belonging to a onefold excitation operators $X_I^A, I \in \mathrm{occ}, A \in \mathrm{virt}$, will be denoted as $\mu = \binom{I}{A}$.

(ii) For an index $I \in \mathrm{occ}$, let $|I|$ label its position $p \in N]$ in the reference determinant (3.2)and denote $\sigma_I = (-1)^{|I|}$.

(iii) For $\mu \in \mathcal{M}$, we denote

$$\rho_\mu := \frac{1}{r(\mu)-1}. \tag{3.16}$$

(iv) Finally, for each $\mu \in \mathcal{M}$, we define a corresponding mapping $\mu : \mathrm{occ} \to \mathcal{I}$: If $I \notin \mu$ (i.e. if the occupied orbital χ_I is "not excited by X_μ"), we let $\mu(I) = I$; if $I \in \mu$, we have in equation (3.7) that $I = I_s$ for some $s \in r]$, and I_s defines by the ordering on $\mathcal{I}$ a unique virtual index A_s (to which the orbital χ_I is "excited by X_μ"), for which we then define $\mu(I) = A_s$.

The first estimate in next lemma was already proven in [190], where it was central to the analysis for the projected CC equations the discrete setting. We re-formulate it here with an improved constant and derive from it the estimate (3.18), which will be useful to show continuity of $T^\dagger$.

Lemma 3.9. *(Estimate for the proof of Theorem 3.6)*

For any sequences $(d_\beta)_{\beta\in\mathcal{M}}, (e_\beta)_{\beta\in\mathcal{M}} \in \ell_2(\mathcal{M})$, there holds

$$\sum_{\nu\in\mathcal{M}} \Big| \sum_{\beta\in\mathcal{M}} d_\beta e_{\nu\ominus\beta} \Big|^2 \;\le\; C_N \, \|(d_\beta)_{\beta\in\mathcal{M}}\|_{\ell_2(\mathcal{M})}^2 \, \|(e_\beta)_{\beta\in\mathcal{M}}\|_{\ell_2(\mathcal{M})}^2 \tag{3.17}$$

and also

$$\sum_{\nu\in\mathcal{M}} \Big| \sum_{\beta\in\mathcal{M}} d_\beta e_{\nu\oplus\beta} \Big|^2 \;\le\; C_N \, \|(d_\beta)_{\beta\in\mathcal{M}}\|_{\ell_2(\mathcal{M})}^2 \, \|(e_\beta)_{\beta\in\mathcal{M}}\|_{\ell_2(\mathcal{M})}^2. \tag{3.18}$$

Proof. We start by estimating the number of indices μ for which $\mu \subseteq \nu$ holds for a fixed index ν (and thus for the number of indices μ for which $\nu \ominus \mu$ gives a nonzero contribution): By definition, $\mu \subseteq \nu$ iff $\mathrm{virt}(\mu) \subseteq \mathrm{virt}\,(\nu)$ and $\mathrm{occ}(\nu) \subseteq \mathrm{occ}\,(\mu)$, so the number of possible indices $\mu \subseteq \nu$ for which Φ_μ has excitation rank s is given by $\binom{r}{s}\binom{N}{(N-s)-(N-r)} = \binom{r}{s}\binom{N}{r-s}$, where r denotes the excitation rank of Φ_ν. Summing up over all ranks $s \leq r$ gives

$$\sum_{0\leq s\leq r} \binom{r}{s}\binom{N}{r-s} \;=\; \binom{N+r}{r} \;\leq\; \binom{2N}{N} \;=:\; C_N$$

by Vandermonde's identity and a (sharp) worst-case estimate. Now, we can estimate the left hand of (3.17) by noting that for every fixed ν, the sum over β contains at most C_N non-null summands; thus

$$\sum_{\nu\in\mathcal{M}} \big|\sum_{\beta\in\mathcal{M}} d_\beta e_{\nu\ominus\beta}\big|^2 \;\leq\; C_N \sum_{\nu\in\mathcal{M}}\sum_{\beta\in M} |d_\beta|^2 |e_{\nu\ominus\beta}|^2 \;\leq\; C_N \sum_{\beta\in\mathcal{M}} |d_\beta|^2 \sum_{\nu\in\mathcal{M}} |e_\nu|^2,$$

giving (3.17).
To prove (3.18), we note that (3.17) means that for $(d_\beta)_{\beta\in\mathcal{M}} \in \ell_2(\mathcal{M})$, the mapping

$$M : (f_\delta)_{\delta\in\mathcal{M}} \mapsto \big(\sum_{\nu\in\mathcal{M}} f_\nu d_{\delta\ominus\nu}\big)_{\delta\in\mathcal{M}}$$

is a continuous mapping $\ell_2(\mathcal{M}) \to \ell_2(\mathcal{M})$ with continuity constant $\|M\| \leq C_N^{\frac{1}{2}} \|d_\beta\|_{\ell_2}$. We compute the adjoint of M: Because there holds for $(e_\delta)_\delta \in \ell_2(\mathcal{M})$ that

$$\big\langle M(f_\delta)_{\delta\in\mathcal{M}}, (e_\delta)_{\delta\in\mathcal{M}}\big\rangle \;=\; \sum_{\delta\in\mathcal{M}}\sum_{\nu\in\mathcal{M}} f_\nu d_{\delta\ominus\nu} e_\delta \;=\; \big\langle (f_\nu)_{\nu\in\mathcal{M}}, (\sum_{\delta\in\mathcal{M}} d_{\delta\ominus\nu} e_\delta)_{\nu\in\mathcal{M}}\big\rangle$$

and for fixed $\nu \in \mathcal{M}$ that

$$\sum_{\delta\in\mathcal{M}} d_{\delta\ominus\nu} e_\delta \;=\; \sum_{\nu\subseteq\delta\in\mathcal{M}} d_{\delta\ominus\nu} e_\delta \;=\; \sum_{\beta\in\mathcal{M}} d_\beta e_{\nu\oplus\beta},$$

$M^\dagger$ is given by

$$M^\dagger : \;\; (e_\beta)_{\beta\in\mathcal{M}} \;\mapsto\; \big(\sum_{\beta\in\mathcal{M}} d_\beta e_{\nu\oplus\beta}\big)_{\nu\in\mathcal{M}}.$$

$M^\dagger$ is continuous with $\|M^\dagger\| \leq C_N^{\frac{1}{2}} \|(d_\beta)_{\beta\in\mathcal{M}}\|_{\ell_2}$, and writing this out gives (3.18).

□

Using the estimates (3.17), the proof of the L^2-continuity of T is completely analogous to the proof of [190], Lemma 4.13, for the discrete case. We therefore leave it out for sake of brevity. To prove the continuity of $T : H^1 \to H^1$, we now equip H^1 with the equivalent norm induced by the preconditioning mapping F. The following lemma provides a working expression for the F-norm of a wave function Ψ.

Lemma 3.10. *(F-norm of antisymmetric functions)*
Let $\overline{\chi}_P := F^{-\frac{1}{2}}\chi_P$ *for all* $P \in \mathcal{I}$. *For any* $\Psi = \sum_{\mu\in\mathcal{M}} d_\mu \Psi_\mu \in \mathbb{H}^1$, *there holds*

$$\|\Psi\|_F^2 = \sum_{J\in\mathrm{occ}} \sum_{\nu\in\mathcal{M}} \Big| \sum_{\substack{I\in\mathrm{occ}\\ I\notin\nu}} \sigma_I \; d_\nu \; \langle \chi_I, \overline{\chi}_J\rangle_F \Big|^2 \tag{3.19}$$

$$+ \sum_{B\in\mathrm{virt}} \sum_{\nu\in\mathcal{M}} \rho_\nu \Big| \sum_{I\in\mathrm{occ}} \sum_{A\in\mathrm{virt}} \sigma_I d_{\nu\oplus\binom{A}{I}} \langle \chi_A, \overline{\chi}_B\rangle_F \Big|^2. \tag{3.20}$$

Proof. We will show that for any $i \in N]$, there holds

$$\|\Psi\|_{\hat{F}_i}^2 = \frac{1}{N}\Big(\sum_{\substack{J\in\mathrm{occ}\\ \nu\in\mathcal{M}}} \Big| \sum_{\substack{I\in\mathrm{occ}\\ I\notin\nu}} \sigma_I \; d_\nu \; \langle \chi_I, \overline{\chi}_J\rangle_F \Big|^2 + \sum_{\substack{B\in\mathrm{virt}\\ \nu\in\mathcal{M}}} \rho_\nu \Big| \sum_{\substack{I\in\mathrm{occ}\\ A\in\mathrm{virt}}} \sigma_I d_{\nu\oplus\binom{A}{I}} \langle \chi_A, \overline{\chi}_B\rangle_F \Big|^2 \Big). \tag{3.21}$$

By definition of $F = F_N$ (see Def. 1.1), we have $\|\Psi\|_F^2 = \sum_{i=1}^N \|\Psi\|_{\hat{F}_i}^2$ for any $\Psi \in \mathbb{H}^1$, and the lemma is then proven. To make notations not more complicated than necessary, we suppose $i = 1$ without loss of generality. We define an orthonormal basis with respect to the $\hat{F}_1$-inner product: Let us denote by $\overline{\mathcal{M}} \subseteq I^{N-1}$ the set of ordered indices of length $N-1$, and denote for $\overline{\nu} \in \overline{\mathcal{M}}$ by $\Phi_{\overline{\nu}}$ the $(N-1)$-electron Slater determinant formed from the one-particle basis functions (taken from (3.1)) determined by $\overline{\nu}$. Because $\overline{\chi}_P := F^{-\frac{1}{2}}\chi_P$ as defined above is a F-orthonormal one-particle spin basis, the set

$$\overline{\mathbb{B}} \;:=\; \{\Psi_{P\overline{\nu}} := \overline{\chi}_P \otimes \Phi_{\overline{\nu}} \;\mid\; P \in \mathcal{J}, \overline{\nu} \in \overline{\mathcal{M}}\}$$

is an $\hat{F}_1$-orthonormal system. We can write every basis function $\Psi_\mu \in \mathbb{B}$ as

$$\Psi_\mu \;=\; \frac{1}{N!} \sum_{\pi\in S(N)} (-1)^{|\pi|} \; \chi_{\mu_{\pi(1)}} \otimes \ldots \otimes \chi_{\mu_{\pi(N)}} \;=\; \frac{1}{N} \sum_{I\in\mathrm{occ}} \sigma_I \; \chi_{\mu(I)} \otimes \Phi_{\overline{\mu}_I}, \tag{3.22}$$

where $\Phi_{\overline{\mu}_I}$ is the Slater determinant from $\overline{\mathbb{B}}$ obtained from Ψ_μ by removing the function $\chi_{\mu(I)}$. Therefore, $\mathbb{H}^1$ is contained in the $\hat{F}_1$-span of $\overline{\mathbb{B}}$, and we can calculate the $\hat{F}_1$-norm of any $\Psi \in \mathbb{H}^1$ by expanding Ψ in the basis $\overline{\mathbb{B}}$. To do so, we decompose for fixed $I \in \mathrm{occ}$ the set $\mathcal{M}$ into indices belonging to excitation operators that do not/do contain the annihilator for I,

$$\sum_{\mu\in\mathcal{M}} d_\mu \; (\chi_{\mu(I)} \otimes \Phi_{\overline{\mu}_I}) \;=\; \sum_{\substack{\mu\in\mathcal{M}\\ I\notin\mu}} d_\mu \; (\chi_I \otimes \Phi_{\overline{\mu}_I}) \;+\; \sum_{\substack{\mu\in\mathcal{M}\\ I\notin\mu}} \rho_\mu \sum_{A\in\mathrm{virt}} d_{\mu\oplus\binom{A}{I}} \; (\chi_A \otimes \Phi_{\overline{\mu}_I}).$$

Note that in the second term, there are $r(\mu)+1$ combinations of indices $\mu, \binom{A}{I}$ that give rise to the same summand indexed by $\mu \oplus \binom{A}{I}$, causing the factor ρ_μ. Inserting (3.22) into $\Psi = \sum_{\mu\in\mathcal{M}} d_\mu \Psi_\mu$, interchanging sums and then using the above decomposition gives

$$\Psi = \frac{1}{N} \sum_{I\in\mathrm{occ}} \sigma_I \sum_{\substack{\mu\in\mathcal{M}\\ I\notin\mu}} \Big(d_\mu \; (\chi_{\mu(I)} \otimes \Phi_{\overline{\mu}_I}) + \rho_\mu \sum_{A\in\mathrm{virt}} d_{\mu\oplus\binom{A}{I}} \; (\chi_A \otimes \Phi_{\overline{\mu}_I}) \Big). \tag{3.23}$$

Let $I \in \mathrm{occ}$ and $\overline{\nu} = (I_{\overline{\nu}_1}, \ldots, I_{\overline{\nu}_m}, A_{\overline{\nu}_1}, \ldots, A_{\overline{\nu}_{N-1-m}}) \in \overline{\mathcal{M}}$ be fixed. Then $\overline{\nu}$ defines a unique excitation operator $\nu_I \in \mathcal{M}$ by defining $\mathrm{occ}(\nu_I) = \mathrm{occ} \setminus \{I, I_{\overline{\nu}_1}, \ldots, I_{\overline{\nu}_m}\}$, $\mathrm{virt}(\nu_I) = \{A_{\overline{\nu}_1}, \ldots, A_{\overline{\nu}_{N-1-m}}\}$. The relation $(\overline{\nu}, \nu_I)$ defines a bijection between the set $\overline{\mathcal{M}}$ and the set$\{\mu \in \mathcal{M} | I \notin \mu\}$. If we let $\delta^I_{\overline{\nu},\mu} = 1$ if $\nu_I = \mu$ and zero elsewise, testing (3.23) with $\Psi_{P\overline{\nu}}$ yields

$$\langle \Psi, \Psi_{P\overline{\nu}} \rangle = \frac{1}{N} \sum_{I \in \mathrm{occ}} \sigma_I \sum_{\substack{\mu \in \mathcal{M} \\ I \notin \mu}} \Big(d_\mu \ \langle \chi_I, \overline{\chi}_P \rangle_F \ \delta^I_{\overline{\nu},\mu} + \rho_\mu \sum_{A \in \mathrm{virt}} d_{\mu \oplus \binom{I}{A}} \ \langle \chi_A, \overline{\chi}_P \rangle_F \delta^I_{\overline{\nu},\mu} \Big).$$

Therefore, we get

$$\begin{aligned} \|\Psi\|^2_{\hat{F}_1} &= \frac{1}{N} \sum_{P \in \mathcal{J}} \sum_{\overline{\nu} \in \overline{\mathcal{M}}} \Big| \sum_{I \in \mathrm{occ}} \sigma_I \sum_{\substack{\mu \in \mathcal{M} \\ I \notin \mu}} \Big(d_\mu \ \langle \chi_I, \overline{\chi}_P \rangle_F \ \delta^I_{\overline{\nu},\mu} + \rho_\mu \sum_{A \in \mathrm{virt}} d_{\mu \oplus \binom{I}{A}} \ \langle \chi_A, \overline{\chi}_P \rangle_F \delta^I_{\overline{\nu},\mu} \Big) \Big|^2 \\ &= \frac{1}{N} \sum_{P \in \mathcal{J}} \sum_{\nu \in \mathcal{M}} \Big| \sum_{\substack{I \in \mathrm{occ} \\ I \notin \nu}} \sigma_I \Big(\ d_\mu \langle \chi_I, \overline{\chi}_P \rangle_F \ + \ \rho_\mu \sum_{A \in \mathrm{virt}} d_{\mu \oplus \binom{I}{A}} \langle \chi_A, \overline{\chi}_P \rangle_F \ \Big) \Big|^2. \end{aligned}$$

Using that $d_{\mu \oplus \binom{I}{A}} = 0$ if $I \in \nu$ and the orthogonality condition (3.4), one obtains the desired expression (3.21), implying (3.19f.).

□

Proof of Theorem 3.6: The H^1-continuity of T and $T^\dagger$.
We are now in the position to show that T continuously maps $\mathbb{H}^1 \to \mathbb{H}^1$. We denote

$$\Psi = \sum_{\mu \in \mathcal{M}} c_\mu \Psi_\mu, \quad \Psi^* = \sum_{\alpha \in \mathcal{M}^*} t_\alpha \Psi_\alpha, \quad T\Psi = \sum_{\nu \in \mathcal{M}^*} d_\nu \Psi_\nu = \sum_{\mu \in \mathcal{M}} \sum_{\alpha \in \mathcal{M}^*} t_\alpha c_\mu X_{\alpha \oplus \mu} \Psi_0.$$

We now compute the $\hat{F}$-norm for $T\Psi$ according to Lemma 3.10: For $\nu \in \mathcal{M}, A \in \mathrm{virt}$, there holds

$$\begin{aligned} d_\nu &= \sum_{\mu \in \mathcal{M}} \sum_{\alpha \in \mathcal{M}^*} t_\alpha c_\mu \delta_{\alpha \oplus \mu, \nu} = \sum_{\alpha \in \mathcal{M}^*} t_\alpha c_{\nu \ominus \alpha}, \\ \sum_{I \in \mathrm{occ}} d_{\nu \oplus \binom{A}{I}} &= \sum_{I \in \mathrm{occ}} \sum_{\mu \in \mathcal{M}} \sum_{\alpha \in \mathcal{M}^*} \big(t_{\alpha \oplus \binom{A}{I}} c_\mu \delta_{\alpha \oplus \mu, \nu} + t_\alpha c_{\mu \oplus \binom{A}{I}} \big) \delta_{\alpha \oplus \mu, \nu} \\ &= \sum_{\substack{I \in \mathrm{occ} \\ I \notin \nu}} \sum_{\alpha \in \mathcal{M}^*} t_{\alpha \oplus \binom{A}{I}} c_{\nu \ominus \alpha} + t_{\nu \ominus \alpha} c_{\alpha \oplus \binom{A}{I}}. \end{aligned}$$

Thus, inserting this in (3.19f.),

$$\begin{aligned} \|T\Psi\|^2_F &= \sum_{J \in \mathrm{occ}} \sum_{\nu \in \mathcal{M}} \Big| \sum_{\substack{I \in \mathrm{occ} \\ I \notin \nu}} \sigma_I \sum_{\alpha \in \mathcal{M}^*} t_\alpha c_{\nu \ominus \alpha} \ \langle \chi_I, \overline{\chi}_J \rangle_F \Big|^2 \qquad (3.24) \\ &+ \sum_{B \in \mathrm{virt}} \sum_{\nu \in \mathcal{M}} \rho_\nu \Big| \sum_{I \in \mathrm{occ}} \sum_{A \in \mathrm{virt}} \sigma_I \sum_{\alpha \in \mathcal{M}^*} \big(t_{\alpha \oplus \binom{A}{I}} c_{\nu \ominus \alpha} + t_{\nu \ominus \alpha} c_{\alpha \oplus \binom{A}{I}} \big) \langle \chi_A, \overline{\chi}_B \rangle_F \Big|^2. \end{aligned}$$

Denoting the summand in line (3.24) with (I) and the one in the line below with (II), we can use the estimate (3.17) to obtain for (I) that

$$\begin{aligned}
\text{(I)} &\leq \sum_{J\in\mathrm{occ}}\sum_{\nu\in\mathcal{M}}\Big(\sum_{\substack{I\in\mathrm{occ}\\ I\notin\nu}}\Big|\sum_{\alpha\in\mathcal{M}^*} t_\alpha c_{\nu\ominus\alpha}\,\langle\chi_I,\overline{\chi}_J\rangle_F\Big|\Big)^2\\
&\leq N\cdot\Big(\sum_{I\in\mathrm{occ}}\sum_{J\in\mathrm{occ}}\big|\langle\chi_I,\overline{\chi}_J\rangle_F\big|^2\Big)\sum_{\nu\in\mathcal{M}}\Big|\sum_{\alpha\in\mathcal{M}^*} t_\alpha c_{\nu\ominus\alpha}\Big|^2\\
&\leq NC_N\Big(\sum_{I\in\mathrm{occ}}\|\chi_I\|_F^2\Big)\,\|t_\alpha\|_{\ell_2(\mathcal{M})}^2\,\|c_\alpha\|_{\ell_2(\mathcal{M})}^2\\
&\lesssim \|\Psi^*\|\cdot\|\Psi\|,
\end{aligned}$$

while for (II),

$$\begin{aligned}
\text{(II)}\leq\quad & 2\sum_{B\in\mathrm{virt}}\sum_{\nu\in\mathcal{M}}\rho_\nu\Big|\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I\sum_{\alpha\in\mathcal{M}^*} t_{\alpha\oplus\binom{A}{I}}c_{\nu\ominus\alpha}\langle\chi_A,\overline{\chi}_B\rangle_F\Big|^2 && (3.25)\\
+\ & 2\sum_{B\in\mathrm{virt}}\sum_{\nu\in\mathcal{M}}\rho_\nu\Big|\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I\sum_{\alpha\in\mathcal{M}^*} t_{\nu\ominus\mu}c_{\mu\oplus\binom{A}{I}}\langle\chi_A,\overline{\chi}_B\rangle_F\Big|^2. && (3.26)
\end{aligned}$$

To estimate the summand in line (3.25), we use that for $\alpha\subseteq\nu$, $\rho_\nu\leq\rho_\alpha$, and apply (3.17) afterwards to obtain

$$\begin{aligned}
&\sum_{B\in\mathrm{virt}}\sum_{\nu\in\mathcal{M}}\rho_\nu\Big|\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I\sum_{\alpha\in\mathcal{M}^*} t_{\alpha\oplus\binom{A}{I}}c_{\nu\ominus\alpha}\langle\chi_A,\overline{\chi}_B\rangle_F\Big|^2\\
\leq\ &\sum_{B\in\mathrm{virt}}\sum_{\nu\in\mathcal{M}}\Big|\sum_{\alpha\in\mathcal{M}^*}\Big(\rho_\alpha\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I\, t_{\alpha\oplus\binom{A}{I}}\langle\chi_A,\overline{\chi}_B\rangle_F\Big)c_{\nu\ominus\alpha}\Big|^2\\
\lesssim\ &\Big(\sum_{B\in\mathrm{virt}}\sum_{\alpha\in\mathcal{M}}\rho_\alpha\Big|\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I t_{\alpha\oplus\binom{A}{I}}\langle\chi_A,\overline{\chi}_B\rangle_F\Big|^2\Big)\cdot\|(c_\nu)_{\nu\in\mathcal{M}}\|_{\ell_2(\mathcal{M})}\\
\leq\ &\|\Psi^*\|_F\cdot\|\Psi\|
\end{aligned}$$

by comparison with the expression for the F-norm of Ψ^*, while the same proceeding with the summand in line (3.26) gives the other way around

$$2\sum_{B\in\mathrm{virt}}\sum_{\nu\in\mathcal{M}}\rho_\nu\Big|\sum_{I\in\mathrm{occ}}\sum_{A\in\mathrm{virt}}\sigma_I\sum_{\alpha\in\mathcal{M}^*} t_{\nu\ominus\mu}c_{\mu\oplus\binom{A}{I}}\langle\chi_A,\overline{\chi}_B\rangle_F\Big|^2\ \lesssim\ \|\Psi^*\|\cdot\|\Psi\|_F.$$

Thus altogether, $\|T\Psi\|_F\lesssim\|\Psi^*\|_F\cdot\|\Psi\|_F$, and observing $\|T\Psi_0\|=\|\Psi^*\|$ finishes the first part of the proof. It remains to show the $\mathbb{H}^1$-continuity of $T^\dagger$, for which the proof is analogous to that for T, with the estimate (3.18) entering instead of (3.17); we therefore only sketch the proceeding. Again, the representation (3.19f.) is used to compute $\|T^\dagger\Psi\|_F$. Denoting

$$T^\dagger\Psi\ =\ \sum_{\nu\in\mathcal{M}} d_\nu\Psi_\nu\ =\ \sum_{\alpha\in\mathcal{M}^*}\sum_{\mu\in\mathcal{M}} t_\alpha c_\mu X_{\mu\ominus\alpha}\Psi_0,$$

the coefficients d_n are this time for fixed $I\in\mathcal{I}$, $\nu\in\mathcal{M}$, $I\notin\nu$ given by

$$d_\nu=\sum_{\alpha\in\mathcal{M}^*} t_\alpha c_{\nu\oplus\alpha};\qquad d_{\nu\oplus\binom{A}{I}}\ =\ \sum_{\alpha\in\mathcal{M}^*} t_\alpha c_{\nu\oplus\alpha\oplus\binom{A}{I}}.$$

Inserting this in (3.19f.) for $\|T^\dagger\Psi\|_F$ gives two terms, which can be estimated analogously to the above, only that $\rho_{\nu\oplus\alpha} \leq (N+1)\rho_\nu$ enters instead of $\rho_\alpha \leq \rho_\nu$. We then obtain

$$\|T^\dagger\Psi\|_F \;\lesssim\; \|\Psi^*\| \cdot \|\Psi\| \;+\; \|\Psi^*\| \cdot \|\Psi\|_F \;\lesssim\; \|\Psi^*\| \cdot \|\Psi\|_F,$$

and thus the upper bound for the $\mathbb{H}^1$-norm of $T^\dagger$.

□

Note that the F-norm of Ψ^* does not enter the above estimate. Therefore, the H^1-norm of $T = T_{\Psi^*}$ is not uniformly bounded from below by the H^1-norm of Ψ^* because we can choose a sequence Ψ_n^* for which $\|\Psi_n^*\|_F = 1$ but $\|\Psi_n^*\| \to 0$; there then holds $\|T^\dagger_{\Psi_n^*}\|_F / \|\Psi_n^*\|_F \leq \|\Psi_n^*\| / \|\Psi^*\|_F \to 0$.

Corollary 3.11. *(Continuity of $T : H^{-1} \to H^{-1}$)*

Each cluster operator $T = T_{\Psi^*}$, $\Psi^* \in H^1$, can be extended to a continuous operator $T : H^{-1} \to H^{-1}$. In particular, each excitation operator X_μ can be continuously extended to an operator $H^{-1} \to H^{-1}$, and there holds $T = \sum_{\mu\in\mathcal{M}_k^*} c_\mu X_\mu$ in H^{-1}.

Proof. Because $T^\dagger$ is bounded on H^1, its adjoint $\tilde{T} : H^{-1} \to H^{-1}$ is also continuous with $\|\tilde{T}\|_{H^{-1}\to H^{-1}} = \|T^\dagger\|_{H^1\to H^1}$, and for every $F(\cdot) \in (L^2)' \subseteq H^{-1}$ (which we can write as $\langle\Psi, \cdot\rangle$ with $\Psi \in L^2$), there holds

$$\tilde{T}F \;:=\; F(T^\dagger\cdot) \;=\; \langle\Psi, T^\dagger\cdot\rangle \;=\; \langle T\Psi, \cdot\rangle,$$

so that $\tilde{T}$ defines a continuous extension of T (which we also denoted as T above). Theorem 3.6 in particular implies that $X_\mu : H^{-1} \to H^{-1}$ is continuous and well-defined, and T and $\sum_\mu c_\mu X_\mu$ coincide on the dense subset L^2, so $T = \sum_{\mu\in\mathcal{M}_k^*} c_\mu X_\mu$ also follows.

□

(ii) The linked and the unlinked Coupled Cluster equations. We are now in the position to define the continuous version of the Coupled Cluster equations. With the previous results, the eigenvalue equation (3.5) can be rewritten in terms of the cluster operator T as the problem of finding a coefficient vector $t^* = (t_\alpha)_{\alpha\in\ell_2(\mathcal{M})} \in \ell_2(\mathcal{M}^*)$ such that for $T = \sum_{\alpha\in\mathcal{M}_k^*} t_\alpha X_\alpha$ there holds $\Psi^* := T\Psi_0 \in H^1$ and

$$\langle\Psi_\mu,\; (\hat{H} - E^*)\,(I+T)\Psi_0\rangle \;=\; 0 \quad \text{for all} \;\; \Psi_\mu \;\in\; \mathbb{B}_k;$$

the solution of (3.5) is then given by $\Psi = \Psi_0 + \Psi^*$. The Coupled Cluster method now replaces the above linear parametrisation $I+T$ by an exponential parametrisation. Before we do so, note that in the above, only coefficient vectors $t^* = (t_\alpha)_{\alpha\in\mathcal{M}_k^*}$ are admitted for which the corresponding function Ψ^* is contained in H^1. This is reflected by restricing the set of admissible coefficients from $\ell_2(\mathcal{M}^*)$ in the following way.

Definition 3.12. *(The H^1-coefficient space $\mathbb{V}$)*
Let $\langle \cdot, \cdot \rangle_{\hat{F}} : (\mathrm{span}\{\Psi_0\})^\perp \times (\mathrm{span}\{\Psi_0\})^\perp \to \mathbb{R}$ denote an inner product which on $(\mathrm{span}\{\Psi_0\})^\perp$ induces a norm equivalent to the H^1-norm. We define a subspace $\mathbb{V} \subseteq \ell_2(\mathcal{M}_k^*)$ by

$$\mathbb{V} \;:=\; \{ t \in \ell_2(\mathcal{M}_k^*) \mid \|t\|_{\mathbb{V}} < \infty \}\,. \tag{3.27}$$

where

$$\langle t, s \rangle_{\mathbb{V}} \;:=\; \langle \sum_{\alpha \in \mathcal{M}_k^*} t_\alpha \Psi_\alpha, \sum_{\beta \in \mathcal{M}_k^*} s_\beta \Psi_\beta \rangle_{\hat{F}}^2, \qquad \|t\|_{\mathbb{V}} \;:=\; \langle t, t \rangle_{\mathbb{V}}^{1/2}. \tag{3.28}$$

□

Obviously, the above definition of $\mathbb{V}$ is independent of the particular choice of the norm $\|\cdot\|_{\hat{F}}$. Denoting as $T(t)$ the cluster operator defined by t and $\Psi(t) := T(t)\Psi_0$, there holds

$$\|t\|_{\mathbb{V}} \quad \sim \quad \|\Psi(t)\|_{H^1}; \tag{3.29}$$

in particular, $t \in \mathbb{V}$ iff $\Psi^*(t) \in H^1 \cap (\mathrm{span}\{\Psi_0\})^\perp$, so $(\mathbb{V}, \langle \cdot, \cdot \rangle_{\mathbb{V}})$ is complete and thus is a Hilbert space.

From Theorem 3.6 and (3.29), we infer the following corollary.

Corollary 3.13. *The linear mappings*

$$t \;\mapsto\; T(t) = \sum_{\alpha \in \mathcal{M}^*} t_\alpha X_\alpha, \quad t \;\mapsto\; T^\dagger(t) = \sum_{\alpha \in \mathcal{M}^*} t_\alpha X_\alpha^\dagger$$

are bounded linear mappings $(\mathbb{V}, \|\cdot\|_{\mathbb{V}}) \;\to\; (B(H^1), \|\cdot\|_{H^1 \to H^1})$.

Remark 3.14. (Practical computation of $\|t\|_{\mathbb{V}}$)

Of course, we may use the preconditioning mapping F from (3.3) to induce a norm on $\mathbb{V}$; unfortunately, the shift μ which turns the Fock or Kohn-Sham opeorator into a positive definitite mapping is usually unknown. For practical purposes like error estimation, Lemma 1.26 offers a way out: The lifted Fock operator $F_{HF} = F_{HF,N}$ resp. any Kohn-Sham operator $F_{KS} = F_{KS,N}$ fulfilling the Gårding estimate (1.66), cf. Remark 2.8, fulfils the conditions of Lemma 1.26: therefore, if the respective operator is shifted by its trace $\mathrm{tr}F_{HF}$ resp. $\mathrm{tr}(F_{KS})$, corresponding to the sum Λ_0 of the N eigenvalues belonging to the subspace spanned by the occupied orbitals, $\hat{F} = F - \Lambda_0 I$ (with the computable shift Λ_0) can be used to define and evaluate the norm on $\mathbb{V}$.
Note also that although this mapping $F - \Lambda_0 I$ is particularly convenient to handle if B^Σ is an eigenbasis of the operator F, so that F is diagonal in this basis, evaluation of F in a non-orthogonal, non-eigenbasis may also be performed within reasonable complexity if F is a one-particle operator like F_{HF} or F_{KS}.

□

To formulate the CC equations, we need one more lemma justifying the exponential parametrisation; it is the continuous version of [190], Lemma 4.2, and Theorem 4.3.

Lemma 3.15. *(Properties of the exponential function on the algebra of cluster operators)* *The set* $L := \{t_0 I + T(t) \mid t_0 \in \mathbb{R},\ t \in \mathbb{V}\}$ *is a closed commutative subalgebra of* $B(H^1)$*, containing zero as the only non-invertible element. The exponential function* $\exp(X) = \sum_{i=0}^{N} X^i/i!$ *is a local* C^∞*-diffeomorphism mapping onto* $L\backslash\{0\}$*. In particular,* exp *is a bijection between the sets*

$$\mathcal{T} = \{T(t) \mid t \in \mathbb{V}\} \quad \textit{and} \quad I + \mathcal{T} = \{I + T(t) \mid t \in \mathbb{V}\}.$$

The lemma also holds if H^1 *is replaced by* H^{-1} *in the above, or if* $\mathbb{V}$ *is replaced by a subspace* $\mathbb{V}_d \subseteq \mathbb{V}$*.*

Proof. Taking Theorem 3.6 into account, the proof for the properties of L is identical with that from [190], Lemma 4.2, and Theorem 4.3. Because L is a commutative subalgebra of H^1 resp. H^{-1}, the exponential function is a local C^∞-diffeomorphism on $L\backslash\{0\}$, see e.g. [182]. The series terminates at $i = N$ because any product of more than N excitation operators contains more than N annihilators for the N occupied orbitals and thus has to vanish, see Lemma 1.20(v). exp maps $\mathcal{T}$ to $I + \mathcal{T}$ by definition, and on $I + \mathcal{T}$, its inverse is given by the (terminating) logarithmic series $\log(X) = \sum_{i=1}^{N} (-1)^{i-1}(X - I)^i/i$ (see [190]), which obviously maps to $\mathcal{T}$, so the lemma is proven.

□

We can now show that the exact (weak) eigenproblem (3.5) is equivalent to the continuous Coupled Cluster equations formulated in the following theorem.

Theorem 3.16. *(The continuous Coupled Cluster equations)*

An intermediately normed function $\underline{\Psi} \in H^1$ *together with a corresponding eigenvalue* $E^* \in \mathbb{R}$ *solves the (weak, CI) eigenproblem*

$$\langle \Psi_\mu,\ (\hat{H} - E^*)\underline{\Psi} \rangle \ = \ 0, \quad \textit{for all}\ \ \mu \in \mathcal{M}_k \tag{3.30}$$

if and only if $\underline{\Psi} = e^T \Psi_0$ *for some cluster operator* $T = \sum_{\alpha \in \mathcal{M}^*} t_\alpha X_\alpha$ *for which* $\|t_\alpha\|_{\mathbb{V}} < \infty$*, and which fulfils the* unlinked Coupled Cluster equations

$$\langle \Psi_\mu, (\hat{H} - E^*) e^T \Psi_0 \rangle = 0, \quad \textit{for all}\ \ \mu \in \mathcal{M}_k, \tag{3.31}$$

or equivalently, the linked Coupled Cluster equations*,*

$$E^* \ = \ \langle \Psi_0, \hat{H} e^T \Psi_0 \rangle, \qquad \langle \Psi_\mu, e^{-T} \hat{H} e^T \Psi_0 \rangle = 0, \quad \textit{for all}\ \ \mu \in \mathcal{M}_k^*. \tag{3.32}$$

Note that the above equivalence of linked and unlinked formulation does not need to hold anymore if in a discretised setting, based on certain selection criteria, only some of the amplitudes of the discretised basis are used for a computation. In this case, $e^{T^\dagger}$ is not necessarily surjective anymore; to guarantee this, the set of selected amplitudes has to be excitation complete, which is for instance the case for canonical models like CCSD, CCSDT etc., see [190] for details.

Proof. Using Theorem 3.6, $\Psi \in H^1$ solves the set of equations (3.30) iff there is a continuous cluster operator $S : H^1 \to H^1$ such that $\Psi = (I + S)\Psi_0$ and

$$\langle \Psi_\mu, (\hat{H} - E^*)(I + S)\Psi_0 \rangle = 0, \quad \text{for all } \mu \in \mathcal{M}_k \tag{3.33}$$

By Lemma 3.15, there is a unique cluster operator T such that $I + S = e^T$, so that (3.33) is equivalent to finding $T : H^1 \to H^1$ such that

$$\langle \Psi_\mu, (\hat{H} - E^*)e^T\Psi_0 \rangle = 0, \quad \text{for all } \mu \in \mathcal{M}_k, \tag{3.34}$$

or in other words, $0 = (\hat{H} - E^*)e^T\Psi_0 \in H^{-1}$. By Theorem 3.6, the L_2-adjoint $T^\dagger$ of T is continuous as mapping $H^1 \to H^1$; therefore, $e^{T^\dagger}$ is a continuous invertible mapping $H^1 \to H^1$, and (3.34) is equivalent to

$$\langle e^{-T^\dagger}\Psi, (\hat{H} - E^*)e^T\Psi_0 \rangle = 0, \quad \text{for all } \Psi \in H^1. \tag{3.35}$$

Due to the continuity of the adjoint mapping $A \mapsto A^\dagger$, we have

$$\langle e^{-T^\dagger}\Psi, (\hat{H} - E^*)e^T\Psi_0 \rangle \;=\; \langle \Psi, (e^{-T^\dagger})^\dagger(\hat{H} - E^*)e^T\Psi_0 \rangle \;=\; \langle \Psi, e^{-T}(\hat{H} - E^*)e^T\Psi_0 \rangle$$

with the exponential e^{-T} of $-T$ taken in H^{-1}, completing the proof.

□

3.3 Analytical properties of the Coupled Cluster function

The linked Coupled Cluster equations (3.32) exhibit certain advantages when it comes to implementation (see Section 3.5), and are therefore almost exclusively used in practice. For this reason, we will now concentrate on the analysis of the linked Coupled Cluster equations and rewrite them as a root problem for the so-called Coupled Cluster function.[34] We will then show that the CC function is locally Lipschitz continuous and locally strongly monotone; these results will then be used to obtain results for existence and local uniqueness in Theorem 3.21 and a local "goal oriented" error estimator in Theorem 3.24. First, we compile the following properties of f, which were already shown in [190] for the discrete case.

Definition/Lemma 3.17. *(The Coupled Cluster function)*

We define the Coupled Cluster function

$$f : \mathbb{V} \;\to\; \mathbb{V}', \quad f(t) \;:=\; \big(\langle \Psi_\alpha, e^{-T}\hat{H}e^{T}\Psi_0 \,\rangle\big)_{\alpha\in\mathcal{M}_k^*}. \tag{3.36}$$

mapping $\mathbb{V}$ to its dual $\mathbb{V}'$ and is C^∞ on $\mathbb{V}$. f and all derivatives $f^{(n)}$ of f are Lipschitz-continuous on bounded domains of $\mathbb{V}$.
An intermediately normed function $\underline{\Psi} = (I + T(t^*))\Psi_0$ is a weak eigenvector of the electronic Schrödinger equation if and only if

$$f(t^*) \;=\; 0 \;\in\; \mathbb{V}'. \tag{3.37}$$

Proof of Lemma 3.17. Let us denote by $\langle \cdot, \cdot\cdot \rangle_{\ell_2}$ the usual $\ell_2(\mathcal{M}_k^*)$-inner product. Then, for $s, t \in \mathbb{V}$, we obtain with the use of (1.39), Theorem 3.6, Corollary 3.13 and Lemma 3.15 that

$$\langle f(t), s\rangle_{\ell_2} \;=\; \sum_{\alpha\in\mathcal{M}_k^*} \langle s_\alpha \Psi_\alpha, e^{-T}\hat{H}e^{T}\Psi_0 \,\rangle \;\le\; \|T(s)\Psi_0\|_{H^1}\|e^{-T}\hat{H}e^{T}\Psi_0\|_{H^{-1}} \;\le\; C(t)\|s\|_{\mathbb{V}},$$

where the constant $C(t)$ depends on the $\mathbb{V}$-norm of t, so that $\langle f(t), \cdot\rangle_{\ell_2}$ defines a continuous functional on $\mathbb{V}$. f is C^∞ as a composition of C^∞-functions. The Lipschitz continuity of f on bounded domains follows from a short calculation based on the same property of $T \mapsto e^T$ as mapping $H^1 \to H^1$ and $H^{-1} \to H^{-1}$, see Lemma 3.15; for the proof that all derivatives of f are Lipschitz-continuous on bounded domains, see [190] which transfers to our case.

□

[34] The Coupled Cluster function is defined on the infinite dimensional space $\mathbb{V}$, and some readers might therefore prefer the term "(nonlinear) operator" and denote it by a capital letter. To keep consistent with quantum chemistry literature and to avoid confusion with the "Fock operator-like" mapping F, we will stick to the physicist's/chemist's nomenclature of "the Coupled Cluster function" f here.

Theorem 3.18. *(Local strong monotonicity of the CC function)*

Let E^* be a simple eigenvalue of H. If the reference determinant Ψ_0 lies in a suitable neighbourhood of the (intermediately normed) solution $\underline{\Psi}$ of the Schrödinger equation, the Coupled Cluster function f is strongly monotone in a neighbourhood of its solution $t^* = (t^*_\alpha)_{\alpha\in\mathcal{M}^*}$, i.e. there are constants $\gamma, \delta > 0$ such that

$$\langle f(s) - f(t), s - t\rangle \quad \geq \quad \gamma \cdot \|s - t\|^2_{\mathbb{V}} \tag{3.38}$$

holds for all $s, t \in \mathbb{V}$ with $\|s - t^*\|_{\mathbb{V}}, \|t - t^*\|_{\mathbb{V}} < \delta$.

The core ingredient to the proof of (3.38) is the following lemma which bases on Lemma 1.23.

Lemma 3.19. *Let $U_0 := span\{\Psi_0\}$. If the reference determinant Ψ_0 lies sufficiently close to the (intermediately normed) solution $\underline{\Psi}$ of the Schrödinger equation and E^* is a simple eigenvalue of $\hat{H}$, the restriction of the mapping $\hat{H} - E^*$ to the orthogonal complement of U_0 is H^1-elliptic, i.e.*

$$\langle \Psi, (\hat{H} - E^*)\Psi\rangle \quad \geq \quad \gamma' \, \|\Psi\|^2_1 \tag{3.39}$$

holds for some $\gamma' > 0$ and all $\Psi \in U_0^\perp$.

Proof. We show ellipticity of $\hat{H} - E^*$ with respect to the L_2-inner product and then apply Lemma 1.23. Let $\underline{P}, P_0$ be the L_2-orthogonal projectors on $\text{span}\{\underline{\Psi}\}$, $\text{span}\{\Psi_0\}$, respectively, and denote the spectral gap by $\gamma^* := \inf(\text{spec}(\hat{H})\backslash\{E^*\}) - E^*$. Because $\hat{H} - E^* = 0$ on $\text{span}\{\underline{\Psi}\}$, there holds for any $\Psi \in U_0^\perp$ that

$$\langle \Psi, (\hat{H} - E^*)\Psi\rangle \;=\; \langle (I - \underline{P})\Psi, (\hat{H} - E^*)(I - \underline{P})\Psi\rangle \;\geq\; \gamma^* \|(I - \underline{P})\Psi\|^2_{L_2}$$

by use of the Courant-Fischer theorem [179]. We want to use $(I - P_0)\Psi = \Psi$, and compute the difference of the projectors: Letting $\bar{\underline{\Psi}} := \underline{\Psi}/\|\underline{\Psi}\|_{L_2}$, a short calculation shows that

$$\|P_0 - \underline{P}\|_{L_2\to L_2} \;=\; \max_{f\in L_2, \|f\|=1} \big|\, \langle f, \Psi_0\rangle\Psi_0 - \langle f, \bar{\underline{\Psi}}\rangle\bar{\underline{\Psi}} \,\big| \;\leq\; \|\Psi_0 - \bar{\underline{\Psi}}\|. \tag{3.40}$$

Using orthogonality of Ψ_0 and $T\Psi_0$, there holds with $\tau = \|T\Psi_0\|, \|\Psi_0\| = 1$ that $\bar{\underline{\Psi}} = (\Psi_0 + T\Psi_0)/(1 + \tau^2)^{1/2}$, and one easily sees by orthogonal decomposition that

$$\|\Psi_0 - \bar{\underline{\Psi}}\|^2 = \big(1 - \frac{1}{(1-\tau^2)^{\frac{1}{2}}}\big)^2 + \frac{\tau^2}{(1-\tau^2)^{\frac{1}{2}}} \;=\; 2\big(\frac{1}{(1+\tau^2)} - \frac{1}{(1+\tau^2)^{\frac{1}{2}}}\big) \;=\; 4\tau^2 + \mathcal{O}(\tau^4).$$

Therefore, we can for instance choose $\tau = \|\Psi_0 - \underline{\Psi}\|$ such that $\|P_0 - \underline{P}\| \leq \frac{1}{2}$, and using $(I - P_0)\Psi = \Psi$ there follows

$$\gamma^*\|(I - \underline{P})\Psi\|^2 \;\geq\; \gamma^*\Big(\, \|(I - P_0)\Psi\| \;-\; \|(P_0 - \underline{P})\Psi\|\Big)^2 \;\geq\; \frac{1}{4}\,\gamma^*\,\|\Psi\|^2.$$

$\hat{H} - E^*$ is thus L_2-elliptic on the complement of U_0. Therefore - because the Hamiltonian fulfils Gårding's inequality (1.66), see (1.39) - Lemma 1.23 implies that there is a constant γ' such that (3.39) holds for all $\Psi \in U_0^\perp$.

□

Proof of Theorem 3.18. To show (3.38), we denote the cluster operator belonging to t^* by $T = \sum_{\alpha \in \mathcal{M}^*} t^*_\nu X_\nu$. We let $g_1 := s - t^*, g_2 := t - t^*$ and write the corresponding cluster operators as G_1, G_2. We expand $e^{G_i}, e^{-G_i}, i = 1, 2$ into a series to obtain

$$e^{-T-G_i}\hat{H}e^{T+G_i}\Psi_0 \;=\; e^{-T}\hat{H}e^{T}\Psi_0 \;-\; G_i e^{-T}\hat{H}e^{T}\Psi_0 \;+\; e^{-T}\hat{H}e^{T}G_i\Psi_0 \;+\; \mathcal{O}(\|g_i\|^2_{\mathbb{V}}).$$

Thus, with $G = G_1 - G_2$,

$$\begin{aligned}
\langle f(s) - f(t), s - t\rangle &= \langle f(t^* + g_1) - f(t^* + g_2), g_1 - g_2\rangle \\
&:= \langle G\Psi_0, e^{-T-G_1}\hat{H}e^{T+G_1}\Psi_0\rangle \;-\; \langle G\Psi_0, e^{-T-G_2}\hat{H}e^{T+G_2}\Psi_0\rangle \\
&\geq \langle G\Psi_0, e^{-T}\hat{H}e^{T}G\Psi_0\rangle \;-\; \langle G^\dagger G\Psi_0, e^{-T}\hat{H}e^{T}\Psi_0\rangle - \mathcal{O}(\|g_i\|^3_{\mathbb{V}}) \;=:\; (*)
\end{aligned}$$

by the Cauchy-Schwarz inequality and Corollary 3.13. We let $g_1 - g_2 = g := (g_\alpha)_{\alpha \in \mathcal{M}^*}$ and now have to show that $(*)$ is bounded from below by $\gamma \cdot \|g\|^2_{\mathbb{V}}$. Computation of $G^\dagger G\Psi_0$ then yields

$$G^\dagger G\Psi_0 = \sum_{\nu,\mu \in \mathcal{M}^*} g_\nu g_\mu X_{\nu \ominus \mu}\Psi_0 \;=\; \sum_{\mu \in \mathcal{M}^*} g_\mu^2 \Psi_0 \;+\; \sum_{\nu \in \mathcal{M}^*} \sum_{\substack{\mu \in \mathcal{M}^* \\ \mu \subsetneq \nu}} g_\nu g_\mu X_{\nu \ominus \mu}\Psi_0.$$

Note that in the second sum, for all possible combinations of ν, μ turning up, $X_{\nu\ominus\mu}\Psi_0$ is a determinant of excitation rank greater than zero. Therefore, using that $e^{-T}\hat{H}e^{T}\Psi_0 = E^*\Psi_0$,

$$\langle G^\dagger G\Psi_0, e^{-T}\hat{H}e^{T}\Psi_0\rangle = \sum_{\mu \in \mathcal{M}^*} g_\mu^2 E^* \langle \Psi_0, \Psi_0\rangle = E^*\|g\|^2_{\ell_2} = E^*\langle G\Psi_0, G\Psi_0\rangle.$$

Thus, $(*)$ coincides up to second order with

$$\langle G\Psi_0, e^{-T}\hat{H}e^{T}G\Psi_0\rangle - E^*\langle G\Psi_0, G\Psi_0\rangle \;=\; \langle G\Psi_0, e^{-T}(\hat{H} - E^*)e^{T}G\Psi_0\rangle,$$

and it suffices to show that this expression is bounded from below by $c \cdot \|g\|^2_{\mathbb{V}}$. We expand e^T, e^{-T} into a power series as above to obtain

$$\begin{aligned}
&\langle G\Psi_0, e^{-T}(\hat{H} - E^*)e^{T}G\Psi_0\rangle \\
=\;& \langle G\Psi_0, (\hat{H} - E^*)G\Psi_0\rangle \;+\; \langle G\Psi_0, (\hat{H} - E^*)(T - T^\dagger)G\Psi_0\rangle \;-\; \mathcal{O}(\|t^*\|^2_{\mathbb{V}}\|g\|^2_{\mathbb{V}}) \\
\geq\;& \gamma' \,\|G\Psi_0\|^2_{H^1} \;-\; (\overline{\Lambda} - E^*)\,\|T - T^\dagger\|_{H^1 \to H^1}\|G\Psi_0\|^2_{H^1} \;-\; \mathcal{O}(\|t^*\|^2_{\mathbb{V}}\|g\|^2_{\mathbb{V}})
\end{aligned}$$

where Lemma 3.19 was used in the last step, and the constant $\overline{\Lambda}$ is an upper bound for the norm of $\hat{H} : H^1 \to H^{-1}$. Using $\|T\|_{H^1 \to H^1}, \|T^\dagger\|_{H^1 \to H^1} \lesssim \|t^*\|_{\mathbb{V}}$, $\|G\Psi_0\|_{H^1} \gtrsim \|g\|_{\mathbb{V}}$ thus gives for $\|t^*\|_{\mathbb{V}}$ small enough constants $c, \gamma'', \gamma > 0$ such that

$$\langle G\Psi_0, e^{-T}(\hat{H} - E^*)e^{T}G\Psi_0\rangle \;\geq\; \gamma''\|g\|^2_{\mathbb{V}} \;-\; c\|t^*\|_{\mathbb{V}}\|g\|^2_{\mathbb{V}} \;\geq\; \gamma\|g\|^2_{\mathbb{V}},$$

and the proof is finished.

Corollary 3.20. *(Properties of the derivative of f)*

Let the assumptions of Theorem 3.18 hold. For $s \in U_\delta(t^*)$, the derivatives $Df(s) \in L(\mathbb{V}, \mathbb{V}')$ of the Coupled Cluster function f at s are uniformly bounded, $\mathbb{V}$-coercive linear operators, i.e. there is a $C > 0$ such that

$$\langle Df(s)u, v\rangle \;\leq\; C \cdot \|u\|_{\mathbb{V}}\|v\|_{\mathbb{V}}, \qquad \langle Df(s)u, u\rangle \;\geq\; \gamma\|u\|_{\mathbb{V}}^2 \tag{3.41}$$

holds for all $s \in U_\delta(t^*)$ and $u, v \in \mathbb{V}$.[35]

Proof. The CC function f is C^∞ by Lemma 3.17, and it was already noted above that $Df(t)$ is locally Lipschitz continuous, implying the uniform boundedness. For the coercivity, we expand f into a Taylor series, $f(s+u') - f(s) = Df(s)u' + \mathcal{O}(\|u'\|_{\mathbb{V}}^2)$. Inserting this into the strong monotonicity estimate (3.38), one obtains by choosing $u' = u/c$ small enough and then using linearity that

$$\langle Df(s)u, u\rangle \quad\geq\quad \gamma\,\|u\|_{\mathbb{V}}^2 \;-\; \mathcal{O}(\|u\|_{\mathbb{V}}^3) \quad\geq\quad \gamma\|u\|_{\mathbb{V}}^2 \;-\; \varepsilon$$

holds for all $u \in \mathbb{V}$ and $\varepsilon > 0$. This completes the proof.

□

[35] In this, γ coincides with the monotonicity constant from Theorem 3.18.

3.4 Existence and uniqueness statements and error estimates

We now use the just proven properties of f to obtain results about existence and (local) uniqueness for solutions of the problem (3.37) and for discretisations thereof. Note that our situation is a little different from what is usually assumed in the theory of standard nonlinear functional analysis [80, 215], where existence and uniqueness of continuous as well as discrete solutions follows if f is globally Lipschitz continuous and globally strongly monotone (i.e. (3.38) holds on all of $\mathbb{V}$), see e.g. [68]. This cannot be true in our case if the eigenvalue problem (3.5) has a second solution, corresponding to a bound state aside from the ground state. Instead, existence of the solution of the continuous problem is in our case guaranteed by Assumption 1.13 together with Lemma 3.17, and we will prove the existence of local solutions of the corresponding discretised equations using some well-known results from operator analysis. Concerning uniqueness of continuous and discrete solutions, local statements are the best we can hope for if there are bound states aside from the ground state, and a result of that kind is given in the following theorem.

Theorem 3.21. *(Existence and uniqueness of solutions; quasi-optimality)*

Let the assumptions of Theorem 3.18 be fulfilled. The Coupled Cluster function then possesses a Lipschitz continuous inverse f^{-1} on $B_\delta(t^*)$; in particular, the solution t^* of the Coupled Cluster function that belongs to the lowest eigenvalue of (3.5) is unique in the neighbourhood $B_\delta(t^*)$.
Let $\mathbb{V}_d$ be a subspace of $\mathbb{V}$ for which $d(t^*, \mathbb{V}_d) := \min_{v\in\mathbb{V}_d} \|v - t^*\|_\mathbb{V}$ is sufficiently small. Then the discretised (projected) problem

$$\langle f(t_d), v_d\rangle \;=\; 0 \quad \textit{for all} \quad v_d \in \mathbb{V}_d \tag{3.42}$$

admits a solution t_d in $B_{\delta,d} := \mathbb{V}_d \cap B_\delta(t^*)$ which is unique on $B_{\delta,d}$ and fulfils the quasi-optimality estimate

$$\|t_d - t^*\|_\mathbb{V} \;\leq\; \frac{L}{\gamma}\, d(t^*, \mathbb{V}_d) \tag{3.43}$$

with L the Lipschitz constant of f on $B_\delta(t^*)$. In particular, if $\mathbb{V}_{(n)}$ is a sequence of subspaces of $\mathbb{V}$ for which $\lim_{n\to\infty} d(t^*, \mathbb{V}_{(n)}) \to 0$, the corresponding solutions $t_{(n)} \in B_{\delta,(n)}$ of (3.42) converge to the continuous solution $t^* \in \mathbb{V}$.

The above result shows that if the constant γ in the lower bound (3.39) is close to zero, corresponding to a small gap between the ground state energy and the second lowest energy level, this may not only lead to deterioration of convergence of an e.g. Newton's method employed for solution of the Coupled Cluster equations as experienced in practice, but also means that the constants in the quasi-optimality estimate (3.43) become bad; also, the results proven may in this case only hold on a very small neighbourhood of t^*,

emphasing from another viewpoint the importance of multi-reference approaches in this situation.

Proof of Theorem 3.21. Equation (3.38) implies that f is one-to-one on $B_\delta(t^*)$ and that for $p, q \in f(B_\delta(t^*))$, there holds for the inverse mapping $f^{-1} : f(B_\delta(t^*)) \to B_\delta(t^*)$ that

$$\gamma \|f^{-1}p - f^{-1}q\|^2 \;\leq\; \langle p-q,\; f^{-1}p - f^{-1}q\rangle \;\leq\; \|p-q\|\;\|f^{-1}p - f^{-1}q\|,$$

so f^{-1} is Lipschitz continuous with Lipschitz constant $1/\gamma$. To prove the existence of solutions for sufficiently well discretised problems, we use the following well-known lemma which bases on the fixed point theorem of Brouwder, see e.g. [68], Lemma 4.2.1 for a proof.

Lemma 3.22. *Let $\|.\|_\#$ be an arbitrary norm on $\mathbb{R}^m$, and $\mathbf{h} : \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function on the closed ball $\overline{B}_{R,\|.\|_\#}(\mathbf{0})$ of radius R around $\mathbf{0} \in \mathbb{R}^m$. If $\langle \mathbf{h}(\mathbf{v}), \mathbf{v}\rangle \geq 0$ holds for all $\mathbf{v} \in \mathbb{R}^m$ with $\|\mathbf{v}\|_\# = R$, there is a $\mathbf{v}^* \in \overline{B}_{R,\|.\|_\#}(\mathbf{0})$ for which $\mathbf{h}(\mathbf{v}^*) = \mathbf{0}$.*

To finish the proof of Theorem 3.21, let us now fix a discretisation $\mathbb{V}_d \subseteq \mathbb{V}$ for which $d := d(t^*, \mathbb{V}_d) \leq \delta \cdot \gamma/(\gamma + L)$. We let $m := \dim \mathbb{V}_d$, $\mathbb{B}_d := \{b_j \in \mathbb{V}, j \in \underline{m}\}$ be an orthonormal basis of $\mathbb{V}_d$ and $t^{\mathrm{opt}} = \arg\min d(t^*, \mathbb{V}_d)$.[36] To apply Lemma 3.22, we define for $\mathbf{v} = (v_j)_{j=1}^m \in \mathbb{R}^m$ that $v = \sum_{j=1}^m v_j b_j$ and $\|\mathbf{v}\|_\# := \|v\|_{\mathbb{V}}$. We let

$$\mathbf{h}(\mathbf{v}) \;=\; \big(\langle f(t^{\mathrm{opt}} + \sum_{j=1}^m v_j\varphi_j),\; \varphi_k\rangle\big)_{k=1}^m$$

and observe that $\mathbf{h}(\mathbf{v}) = \mathbf{0}$ for some $\mathbf{v} \in \overline{B}_{R,\#}(\mathbf{0})$ iff $t^{\mathrm{opt}} + \sum_{j=1}^m v_j b_j \in \overline{B}_{R,\mathbb{V}}(t^{\mathrm{opt}})$ solves the discretised problem (3.42). Choosing $R = \delta - d$, $\overline{B}_{R,\mathbb{V}}(t^{\mathrm{opt}})$ lies in the neighbourhood of t^* where f is strongly monotone, so for all $\mathbf{v} \in \mathbb{R}^m$ with $\|\mathbf{v}\|_\# = R$,

$$\begin{aligned}
\langle \mathbf{h}(\mathbf{v}), \mathbf{v}\rangle \;&:=\; \langle f(t^{\mathrm{opt}} + v), v\rangle \\
&=\; \langle f(t^{\mathrm{opt}} + v) - f(t^{\mathrm{opt}}), v\rangle \;+\; \langle f(t^{\mathrm{opt}}) - f(t^*), v\rangle \\
&\geq\; \gamma\|\mathbf{v}\|_\#^2 - Ld\|\mathbf{v}\|_\# \;=\; R(\gamma R - Ld) \;\geq\; 0,
\end{aligned}$$

and because of the local Lipschitz continuity of f, $\mathbf{h}$ is continuous on $\overline{B}_R(0)$. Thus, $\mathbf{h}$ fulfils the conditions of Lemma 3.22, and if $\mathbf{v}^* = (v_j^*)_{j=1}^m \in \overline{B}_{R,\#}(\mathbf{0})$ solves $\mathbf{h}(\mathbf{v}^*) = \mathbf{0}$, then $t_d := t^{\mathrm{opt}} + \sum_{j=1}^m b_j^*\varphi_j \in \overline{B}_{R,\mathbb{V}}(t^{\mathrm{opt}}) \subseteq B_\delta(t^*)$ is a solution of (3.42). The restriction $\tilde{f} : B_{d,\delta}(t^*) \to \mathbb{V}_d'$ of f is also a strongly monotone function, so that with the same argumentation as for the continuous solution, there can only be one $t_d \in B_{d,\delta}(t^*)$ solving $\tilde{f}(t_d) = 0 \in \mathbb{V}_d'$, proving local uniqueness of the solution. The quasi-optimality estimate and the convergence of the discrete solutions t_d towards the continuous limit t^* now follow from Lipschitz continuity and strong monotonicity of f by standard arguments, see e.g. [215].

□

[36] Note that $\mathbb{V}$ is a Hilbert space, see the remarks after Definition 3.12.

To end this section, we now combine the results just proven with the formalism of goal oriented error estimators developed in [22] and also used in [190] to obtain estimators for the Coupled Cluster energy (3.32),

$$E(t) = \langle \Psi_0, e^{-T(t)} \hat{H} e^{T(t)} \Psi_0 \rangle, \tag{3.44}$$

in terms of the approximation quality of the cluster amplitudes $(t_\alpha)_{\alpha \in \mathcal{M}^*}$ and the corresponding wave functions. To do so, we use that the ground state energy E^* is a minimizer of a corresponding Lagrange functional. Because this Lagrangian is non-symmetric, we cannot expect the error to be quadratic with respect to the error of the wave function as was the case e.g. for the energies obtained by Hartree-Fock or CI, see [134] for a more detailed analysis. Instead, the solution z^* of the *dual problem* (corresponding to the Lagrangian multipliers in the finite-dimensional setting) enters the estimates, and we introduce the necessary terminology in the next lemma.

Lemma 3.23. *(Properties of dual solutions)*

Let $\mathbb{V}_d$ be a sufficiently good subspace of $\mathbb{V}$, and $t^* \in \mathbb{V}$ and $t_d \in \mathbb{V}_d$ solutions the Coupled Cluster equations (3.32) and of the discretised (projected) Coupled Cluster equations respectively,

$$\langle f(t^*), s \rangle = 0 \quad \textit{for all} \quad s \in \mathbb{V}, \qquad \langle f(t_d), s_d \rangle = 0 \quad \textit{for all} \quad s_d \in \mathbb{V}_d. \tag{3.45}$$

Under the assumptions of Theorem 3.18, there is a unique "dual solution" or "Lagrangian multiplier" $z^* \in \mathbb{V}$ determined by t^* such that (t^*, z^*) is a stationary point of the Lagrangian $\mathcal{L}(t, z) = E(t) + \langle f(t), z \rangle$, i.e. (t^*, z^*) solves

$$\mathcal{L}'(t^*, z^*) = \left\{ \begin{array}{c} \langle E'(t^*), s \rangle - \langle Df(t^*)s, z^* \rangle \\ \langle f(t^*), s \rangle \end{array} \right\} = 0 \qquad \textit{for all} \quad s \in \mathbb{V}. \tag{3.46}$$

For a sufficiently good discretisation $\mathbb{V}_d$, there is a corresponding unique $z_d \in \mathbb{V}$ such that (t_d, z_d) solves the discretised equations

$$\mathcal{L}'(t_d, z_d) = \left\{ \begin{array}{c} \langle E'(t_d), s_d \rangle - \langle Df(t_d)s_d, z_d \rangle \\ \langle f(t_d), s_d \rangle \end{array} \right\} = 0 \qquad \textit{for all} \quad s_d \in \mathbb{V}_d \tag{3.47}$$

The discrete dual solution z_d approximates the exact dual solution quasi-optimally in the sense that

$$\|z_d - z^*\|_{\mathbb{V}} \quad \lesssim \quad \Delta \; := \; \max\{d(\mathbb{V}_d, t^*),\ d(\mathbb{V}_d, z^*)\}. \tag{3.48}$$

Proof. By definition, t^* solves the second equation from (3.46), so we only have to show that the first equation $\langle E'(t^*), s\rangle = \langle Df(t^*)s, z^*\rangle$ admits a unique solution z^*. Indeed, this is an equation for the linear operator $Df(t^*)^\dagger : \mathbb{V} \to \mathbb{V}'$, which is bounded and coercive because its adjoint $Df(t^*)$ is by Corollary 3.20. Therefore, the Lax-Milgram theorem (see e.g. [6]) ensures existence and uniqueness of z^*. The same argument holds for z_d if the discretisation is fine enough to guarantee (together with quasi-optimality of t_d) that $Df(t_d)$ is also coercive, cf. Corollary 3.20. To show (3.48), we decompose $z_d - z^* = z_d - \hat{z}_d + \hat{z}_d - z^*$, where $\hat{z}_d \in \mathbb{V}_d$ solves the discrete system

$$\langle E'(t^*), s_d\rangle \quad = \quad \langle Df(t^*)s_d, \hat{z}_d\rangle \quad \text{for all} \quad s_d \in \mathbb{V}_d. \tag{3.49}$$

Because $Df(t^*)$ is a bounded and coercive linear mapping, see Corollary 3.20, $\hat{z}_d$ approximates the solution z^* of the corresponding continuous problem (3.46) quasi-optimally by Cea's lemma [6], $\|\hat{z}_d - z^*\|_{\mathbb{V}} \lesssim d(\mathbb{V}_d, z^*)$. For $\|z_d - \hat{z}\|_{\mathbb{V}}$, we at first note again that $Df(t)$ and also by very similar arguments the derivative $E'(t)$ of the energy expression (3.44) are Lipschitz continuous on bounded neighbourhoods of t^*. We choose $c > 0$ such that by Theorem 3.21, for a each discretisation $\mathbb{V}_d$ for which $d(t^*, \mathbb{V}_d) < c$ there holds $\|t_d - t^*\|_{\mathbb{V}} \quad \leq \quad L/\gamma \; d(t^*, \mathbb{V}_d)$ for the discrete solution t_d, and let $L_{f'}$ and $L_{E'}$ be the Lipschitz constants of $Df(t)$ and $E'(t)$ on $B_{cL/\gamma}(t^*)$. We now obtain using (3.47), (3.49) that

$$\begin{aligned}
\gamma\|z_d - \hat{z}_d\|_{\mathbb{V}}^2 &\leq \langle Df(t_d)(z_d - \hat{z}_d), z_d - \hat{z}_d\rangle \\
&= \langle E'(t_d) - E'(t^*), (z_d - \hat{z}_d)\rangle + \langle (Df(t^*) - Df(t_d))(z_d - \hat{z}_d), \hat{z}_d\rangle \\
&\leq (L_{E'} + L_{f'}\|\hat{z}\|_{\mathbb{V}})\, \|t_d - t^*\|_{\mathbb{V}}\, \|z_d - \hat{z}_d\|_{\mathbb{V}},
\end{aligned}$$

and observe that $\|\hat{z}\|_{\mathbb{V}}$ is bounded by $\|z^*\|_{\mathbb{V}} + c \cdot d(\mathbb{V}_d, z^*)$, so that

$$\|z_d - \hat{z}_d\|_{\mathbb{V}} \quad \lesssim \quad \|t_d - t^*\|_{\mathbb{V}} \quad \lesssim \quad d(\mathbb{V}_d, t^*).$$

Thus, $\|z_d - z^*\|_{\mathbb{V}} \lesssim \Delta$, finishing the proof.

□

The quality of a discrete solution (t_d, z_d) of the above Lagrangian equations can be measured in terms of the primal residual $\rho(t_d)$ and the dual residual $\rho^*(t_d, z_d)$, given by

$$\rho(t_d) \;:=\; \langle f(t_d), \cdot\rangle_{\mathbb{V}} \qquad \rho^*(t_d, z_d) \;:=\; \langle E'(t_d), \cdot\rangle_{\mathbb{V}} - \langle Df(t_d)\cdot, z_d\rangle_{\mathbb{V}} \tag{3.50}$$

The theory developed in [22, 16] now allows to estimate the error of the energy approximation in terms of these primal and dual residuals. We first adapt the original theorem from [16] to our notation in (i) and then derive some quasi-optimality estimates for the Coupled Cluster method in (ii), (iii).

Theorem 3.24. *(Energy estimators)*

(i) (Becker/Rannacher [22], see [16] Proposition 6.2.)

Let $(t^*, z^*) \in \mathbb{V}^2$ and $(t_d, z_d) \in \mathbb{V}_d^2$ be the solutions of minimization problems (3.46), (3.47) for a thrice differentiable functional $\mathcal{L}$. Then there holds

$$E(t^*) - E(t_d) \;=\; \frac{1}{2}\rho(t_d)(z^* - v_d) \;+\; \frac{1}{2}\rho^*(t_d, z_d)(t^* - w_d) \;+\; \mathcal{R}_d^3 \quad (3.51)$$

for all v_d, w_d in $\mathbb{V}_d$, where

$$\mathcal{R}_d^3 = \mathcal{O}(\max\{\|t^* - t_d\|, \|z^* - z_d\|\}^3)$$

depends cubically on the primal and dual errors.

(ii) Let $\mathbb{V}_d$ be a sufficiently large subspace of $\mathbb{V}$ in the sense that for Δ from (3.48), $\Delta < c$ for a suitable $c > 0$, and denote by (t^*, z^*) and (t_d, z_d) the solutions the Coupled Cluster equations and of the discretised (projected) Coupled Cluster equations (3.45), respectively, together with the corresponding unique dual solutions. Under the assumptions of Theorem 3.18, there holds

$$|E(t^*) - E(t_d)| \;\leq\; \|t_d - t^*\|_{\mathbb{V}} \Big(c_1 \, \|t_d - t^*\|_{\mathbb{V}} \;+\; c_2 \, \|z_d - z^*\|_{\mathbb{V}}\Big),$$

$$|E(t^*) - E(t_d)| \;\lesssim\; \big(\, d(\mathbb{V}_d, t^*) \;+\; d(\mathbb{V}_d, z^*) \, \big)^2.$$

where the above constants are specified in the proof.

(iii) Denoting $\underline{\Psi}^{z^*} := \Psi_0 + \Psi^{z^*} := e^{T(z^*)}\Psi_0$, by $\underline{\Psi} = \Psi_0 + \Psi^*$ the solution of the exact eigenproblem (3.5) and by $H^1_{d,\perp}$ the discretisation of $(\mathrm{span}\{\Psi_0\})^\perp$ corresponding to $\mathbb{V}_d$, there holds

$$|E(t^*) - E(t_d)| \;\lesssim\; \| \, e^{T(t_d)}\Psi_0 - \underline{\Psi}\|_{H^1} \cdot \big(\, \| \, e^{T(t_d)}\Psi_0 - \underline{\Psi}\|_{H^1}^2 \;+\; \| \, e^{T(z_d)}\Psi_0 - \underline{\Psi}^{z^*}\|_{H^1} \, \big),$$

$$|E(t^*) - E(t_d)| \;\lesssim\; \Big(\inf_{\Psi \in H^1_{d,\perp}} \|\Psi \;-\; \Psi^*\|_{H^1} \;+\; \inf_{\Psi \in H^1_{d,\perp}} \|\Psi \;-\; \Psi^{z^*}\|_{H^1} \, \Big)^2.$$

Proof. For the proof of (i), cf.[16]. To prove (ii), we choose $\tilde{c} > 0$ such that for a each discretisation $\mathbb{V}_d$ for which $d(t^*, \mathbb{V}_d) < \tilde{c}$, there holds $\|t_d - t^*\|_{\mathbb{V}} \;\leq\; L/\gamma \; d(t^*, \mathbb{V}_d)$ for the discrete solution t_d by Theorem 3.21. We denote by L, $L_{f'}$ and $L_{E'}$ the Lipschitz constants of $f(t)$, $Df(t)$ and $E'(t)$ on $B_{\tilde{c}L/\gamma}(t^*)$, and note that by Corollary 3.20, $\|Df(t)\|_{\mathbb{V}\to\mathbb{V}'}$ is uniformly bounded by a constant C on $B_{\tilde{c}L/\gamma}(t^*)$. We now use (3.46) to rewrite the dual residual by inserting zeros as

$$\rho^*(t_d, z_d)(s) \;=\; \langle E'(t_d) - E'(t^*), s\rangle_{\mathbb{V}} \;+\; \langle \big(Df(t^*) - Df(t_d)\big)s, z^*\rangle_{\mathbb{V}} \;+\; \langle Df(t_d)s, z^* - z_d\rangle_{\mathbb{V}}$$

for arbitrary $s \in \mathbb{V}$. Thus, with (3.51) and the definition of the primal residual $\rho(t_d)$, we obtain that for all v_d, w_d in $\mathbb{V}_d$, there holds according to (i) that

$$\begin{aligned} & 2|E(t^*) - E(t_d)| \\ \leq\ & |\langle f(t_d) - f(t^*), z^* - v_d\rangle_{\mathbb{V}}| \ + \ |\langle E'(t_d) - E'(t^*), t^* - w_d\rangle_{\mathbb{V}}| \\ & + |\ \langle \big(Df(t^*) - Df(t_d)\big)(t^* - w_d), z^*\rangle_{\mathbb{V}}\ | \ + \ |\langle Df(t_d)(t^* - w_d), z^* - z_d\rangle_{\mathbb{V}}| \ + \ 2\mathcal{R}_d^3 \\ \leq\ & L\|t_d - t^*\|_{\mathbb{V}}\|z^* - v_d\|_{\mathbb{V}} \ + \ L_{E'}\|t_d - t^*\|_{\mathbb{V}}\|t^* - w_d\|_{\mathbb{V}} \\ & + L_{f'}\|t^* - t_d\|_{\mathbb{V}}\|t^* - w_d\|_{\mathbb{V}}\|z^*\|_{\mathbb{V}} \ + \ C\ \|t^* - w_d\|_{\mathbb{V}}\|z^* - z_d\|_{\mathbb{V}} \ + \ 2\mathcal{R}_d^3 \ := \ (*). \end{aligned}$$

Inserting $v_d = t_d, w_d = z_d$, we obtain

$$\begin{aligned} |E(t^*) - E(t_d)| \ &\leq \ \frac{1}{2}\|t_d - t^*\|_{\mathbb{V}} \Big((L_{f'}\|z^*\|_{\mathbb{V}} + L_{E'})\ \|t_d - t^*\|_{\mathbb{V}} + (L + C)\ \|z_d - z^*\|_{\mathbb{V}} \Big) + \mathcal{R}_d^3 \\ &=: \quad \|t_d - t^*\|_{\mathbb{V}} \Big(c_1\ \|t_d - t^*\|_{\mathbb{V}} \ + \ c_2\ \|z_d - z^*\|_{\mathbb{V}} \Big) \ + \ \mathcal{R}_d^3. \end{aligned}$$

By Lemma 3.23, $\|z_d - z^*\|_{\mathbb{V}}$ is bounded by Δ; thus, we can (by additionally using the quasi-optimality of t_d, Theorem 3.21) control the remainder term $\mathcal{R}_d^3$ in terms of $\mathcal{O}(\Delta^3)$. Therefore, the first estimate of (ii) is proven by choosing Δ small enough, while the second follows from $(*)$ by inserting for v_d, w_d the best approximations $t^{\mathrm{opt}}, z^{\mathrm{opt}} \in \mathbb{V}_d$ of t^*, z^*. To prove (iii), we utilize Lemma 3.15: The first estimate follows from the first one of (ii) with the observation that locally,

$$\|t - s\|_{\mathbb{V}} \ = \ \|T(t)\Psi_0 - T(s)\Psi_0\|_F \ \sim \ \|e^{T(t)}\Psi_0 - e^{T(s)}\Psi_0\|_F$$

holds; for the second, we use that $\{\exp(T(t)) \mid t \in \mathbb{V}_d\} = \{I + T(t) | t \in \mathbb{V}_d\}$, cf. Lemma 3.15, together with the second estimate given in (ii).

□

3.5 Simplification and evaluation of Coupled Cluster function

(i) Termination of the Baker-Campbell-Haussdorff expansion. We already noted above that for issues of implementation, the linked Coupled Cluster equations (3.32) play a much bigger role than the alternative set (3.31) of equations. This is due to the fact that the term $e^{-T}\hat{H}e^{T}$ can be expanded into the so-called Baker-Campbell-Hausdorff series, which itself terminates because the Hamiltonian is a two-particle operator [103]. Thus, the Coupled Cluster function f can be evaluated exactly within a finite basis set, and each component $(f(t))_\mu, \mu \in \mathcal{M}^*$, contains only products of cluster amplitudes t_α containing at most four factors. We will now give a short and to the author's mind more transparent proof of this fact than the canonical one that can e.g. be found in [103]. After Theorem 3.25 is proven, we will for sake of brevity only make some remarks about how the resulting terms are evaluated in practice and refer the reader to the literature for further reference. To start with, we define for any operator $A : H^1 \to H^{-1}$ the (iterated) commutators $[A,T]_{(0)} := A$, $[A,T]_{(1)} := AT - TA : H^1 \to H^{-1}$ and $[A,T]_{(n)} := [[A,T]_{(n-1)},T]$ for $n \geq 2$, and note that these expressions are well-defined due to Theorem 3.6, Corollary 3.11.

Theorem 3.25. *(Evaluation of the similarity transformed Hamiltonian $e^{-T}\hat{H}e^{T}$)*
For each one-particle operator $\tilde{F} = \sum_{P,Q\in\mathcal{I}} \tilde{f}_{PQ}\, a_P^\dagger a_Q$ and $U = \hat{H} - \tilde{F}$, there holds

$$e^{-T}\hat{H}e^{T} \;=\; \sum_{n=0}^{4}\frac{1}{n!}[\hat{H},T]_{(n)} \;=\; \sum_{n=0}^{2}\frac{1}{n!}[\tilde{F},T]_{(n)} \;+\; \sum_{n=0}^{4}\frac{1}{n!}[U,T]_{(n)}. \qquad (3.52)$$

In the above, the operator $\tilde{F}$ can for instance be the one-particle part in the definition of the Hamiltonian (1.65), or as is often the case in practice, the Fock operator on H^1.
The first part of the proof is the below globalization of the Baker-Campbell-Hausdorff series expansion (for matrices). Afterwards, Lemma 3.28 shows that the iterated commutators $[H,T]_{(n)}$ give zero contributions for $n > 4$.

Lemma 3.26. *For any operator $A : H^1 \to H^{-1}$, there holds the* Baker-Campbell-Hausdorff formula,

$$e^{-T}Ae^{T} \;=\; \sum_{n=0}^{\infty}\frac{1}{n!}[A,T]_{(n)}. \qquad (3.53)$$

Proof. It is not hard to show by induction that $[A,T]_{(n)} = \sum_{i=0}^{n}\binom{n}{i}(-1)^i T^i A T^{n-i}$. Thus,

$$e^{-T}Ae^{T} \;=\; \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{(-T)^i}{i!}A\frac{T^j}{j!} \;=\; \sum_{n=0}^{\infty}\sum_{i=0}^{n}(-1)^i\frac{T^i}{i!}A\frac{T^{n-i}}{(n-i)!} \;=\; \sum_{n=0}^{\infty}\frac{1}{n!}[A,T]_{(n)}.$$

□

Definition 3.27. *(Notations for operator strings)*
Let $\mathcal{E}$ be an arbitrary set of annihilation and creation operators. For a string $B = b_1 \dots b_n$, we denote by

$$\mathcal{C}_{\mathcal{E}}(B) = \Big| \left\{ b_i \mid i \in N \right], \ \exists\, e \ \in \ \mathcal{E} \text{ such that } [b_i, e]_+ \neq 0 \right\} \Big|$$

the number of operators in B that do not anticommute with all contained in $\mathcal{E}$.

□

Lemma 3.28. *Let $\mathcal{E}$ be an anticommuting set of annihilation operators, $[e, f]_+ = 0$ for all $e, f \in \mathcal{E}$, and B, C be operator strings for which $B = b_1 \dots b_{2n}$ contains an even number of operators and for which for $C = c_1 \dots c_m, c_i \in \mathcal{E}$ for all $i \in m]$. Then, if $\mathcal{C}_{\mathcal{E}}(B) = 0$, there holds $[B, C] = 0$, and in case $\mathcal{C}_{\mathcal{E}}(B) \geq 1$, we can write $[B, C] = \sum_{i=1}^{n} B_i$ with operator strings B_i for which $\mathcal{C}_{\mathcal{E}}(B_i) \leq \mathcal{C}_{\mathcal{E}}(B) - 1$.*

Proof. We proceed by induction over m. For $m = 1$, there follows by definition of the anticommutator that by swapping c_1 to the left,

$$\begin{aligned} [B, c_1] &= b_1 \dots b_{2n} c_1 \ - \ c_1 b_1 \dots b_{2n} \\ &= (-1)^{2n} c_1 b_1 \dots b_{2n} \ - \ c_1 b_1 \dots b_{2n} + \sum_{i=0}^{2n} (-1)^i [b_i, c_1]_+ \ b_1 \dots b_{i-1} b_{i+1} \dots b_{2n}. \end{aligned}$$

The first two terms cancel. In the sum in the last, we have in each summand either $[b_i, c_1]_+ = 0$, or that $[b_i, c_1]_+ = 1$ and the operator string $b_1 \dots b_{i-1} b_{i+1} \dots b_{2n}$ contains one operator less not anticommuting with all operators from C. Thus, if $\mathcal{C}_{\mathcal{E}}(B) = 0$, we have $[b_i, c_1]_+ = 0$ for all $1 \leq i \leq N$, so $[B, c_1] = 0$, and if $\mathcal{C}_{\mathcal{E}}(B) \geq 1$, $[B, C]$ is a sum of operator strings B_i for which $\mathcal{C}_{\mathcal{E}}(B_i) \leq \mathcal{C}_{\mathcal{E}}(B) - 1$. For the induction step, we use the same proceeding for $C = c_1 \dots c_{m+1}$ to swap c_{m+1} to the right,

$$\begin{aligned} [B, C] &= b_1 \dots b_{2n} c_1 \dots c_{m+1} \ - \ c_1 \dots c_m c_{m+1} b_1 \dots b_{2n} \\ &= [B, C_m] c_{m+1} + \sum_{i=0}^{2n} (-1)^i [c_{m+1}, b_i]_+ \ c_1 \dots c_m b_1 \dots b_{i-1} b_{i+1} \dots b_{2n}, \end{aligned}$$

where we let $C_m = c_1 \dots c_m$. In the case that $\mathcal{C}_{\mathcal{E}}(B) = 0$, there follows $[B, C_m] c_{m+1} = 0$ by induction hypothesis, and all summands in the second term are also zero because $[c_{m+1}, b_i]_+ = 0$. Thus, $[B, C] = 0$. If $\mathcal{C}_{\mathcal{E}}(B) \neq 0$, we observe for the left term that by induction hypothesis, we can write $[B, C_m]$ as a sum of operator strings B_i for which $\mathcal{C}_{\mathcal{E}}(B_i) \leq \mathcal{C}_{\mathcal{E}}(B) - 1$, so the same holds for $[B, C_m] c_{m+1}$. For the right term, the same argument as in the case $m = 1$ gives that each summand can only contain $\mathcal{C}_{\mathcal{E}}(B) - 1$ operators that do not commute with all operators $c_i \in \mathcal{E}$ (note that the operators from $\mathcal{E}$ anticommute). This completes the proof.

□

Proof of Theorem 3.25. We define

$$\mathcal{E} := \{a_I \mid I \in \mathrm{occ}\} \cup \{a_A^\dagger \mid A \in \mathrm{virt}\}. \tag{3.54}$$

All elements from $\mathcal{E}$ anticommute by Lemma 1.20, and all excitation operators X_α are strings built from elements of $\mathcal{E}$. We write the Hamiltonian $\hat{H}$ as

$$\hat{H} \;=\; \sum_{P,Q} f_{PQ}\; a_P^\dagger a_Q \;+\; \sum_{P,Q,R,S} u_{PQRS}\; a_P^\dagger a_Q^\dagger a_R a_S$$

and obtain

$$\begin{aligned} e^{-T}\hat{H}e^{T} \;=\; \sum_{n=0}^{\infty} \frac{1}{n!}[\hat{H},T]_{(n)} \;=\; & \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{P,Q\in\mathcal{I}} \tilde{f}_{PQ} \quad [a_P^\dagger a_Q, \sum_{\alpha\in\mathcal{M}^*} t_\alpha X_\alpha]_{(n)} \\ & + \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{P,Q,R,S\in\mathcal{I}} u_{PQRS}[a_P^\dagger a_Q^\dagger a_R a_S, \sum_{\alpha\in\mathcal{M}^*} t_\alpha X_\alpha]_{(n)}. \end{aligned}$$

For each string $a_P^\dagger a_Q$, there holds $\mathcal{C}_\mathcal{E}(a_P^\dagger a_Q) \le 2$. We observe that the commutator is linear in its first argument, so iterating Lemma 3.28 gives

$$\mathcal{C}_\mathcal{E}\Big(\big[[a_P^\dagger a_Q, X_{\alpha_1}], X_{\alpha_2}\big]\Big) = 0, \qquad \Big[\big[[a_P^\dagger a_Q, X_{\alpha_1}], X_{\alpha_2}\big], X_{\alpha_3}\Big] = 0$$

for all $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{M}_k^*$, so

$$[a_P^\dagger a_Q, \sum_{\alpha\in\mathcal{M}^*} t_\alpha X_\alpha]_{(n)} \;=\; 0$$

follows for $n \ge 3$. To the iterated commutators $[a_P^\dagger a_Q^\dagger a_R a_S, \sum_{\alpha\in\mathcal{M}^*} t_\alpha X_\alpha]_{(n)}$, an analogous argument applies to show that those of order $n \ge 5$ must give zero contributions, so the proof is finished.

□

Remark 3.29. We remark that the representation (3.53) coincides with the Taylor expansion of the C^∞-function $g(T) = e^{-T}Ae^T$ at $T = 0$: more generally, it is given for arbitrary $S \in B(H^1)$ by $g(S+T) = e^{-T-S}Ae^{T+S} = \sum_{i=0}^4 1/n![e^{-S}Ae^S, T]_{(n)}$. In particular, this implies $f^{(5)} \equiv 0$ for the CC function.

(ii) Evaluation of the iterated commutators. To apply solvers like e.g. the below simple inexact Newton's method to solve the root problem (3.37), the Coupled Cluster function and thus, by Theorem (3.25), the expression

$$f(t) \quad = \quad \sum_{n=0}^{4} \frac{1}{n!} \langle \Psi_\mu, [\hat{H}, T]_{(n)} \Psi_0 \rangle \quad = \quad \sum_{n=0}^{4} \frac{1}{n!} \langle \Psi_0, X_\mu^\dagger [\hat{H}, T]_{(n)} \Psi_0 \rangle \tag{3.55}$$

has to be evaluated. This is a nontrivial task, and for sake of brevity, we only sketch the canonical proceeding here and refer the reader to the comprehensive introduction [56] and the references given therein for deeper insight. To evaluate (3.55), *Wick's theorem*, proven e.g. in [142] and based on the anticommutator relations from Lemma 1.20, is the fundamental tool used to successively reorder the operator strings contained in $X_\mu^\dagger [\hat{H}, T]_{(n)}$ to sums of *normal-ordered* strings (i.e. strings where for $\mathcal{E}$ defined by (3.54), all operators $b \in \mathcal{E}$ are to the left of all operators $b \notin \mathcal{E}$). Normal-ordered strings give a zero contribution to (3.55), and the remainder term is a sum of so called "fully contracted terms", containing no annihilation/creation operators anymore as a result of various anticommutators arising in the ordering process. The task is now to determine all of those admissible fully contracted terms that yield a nonzero distribution to (3.55), and this boils down to choosing the right indices of matrix elements of $\hat{H}$ and of up to four different cluster amplitudes t_ν that contribute to each μ-th component of $f(t)$ according to certain rules. This combinatorial task is nontrivial, and especially tedious to implement because the fully contracted terms feature various combinations of signs arising from the anticommutation laws. Therefore, an approach popularized in [129] is normally used to rewrite the equations as diagrams (see also [56]); then, computable expressions are derived from these diagrams by a fixed set of rules, either by hand or automatedly as e.g. in [12, 104]. To give the reader an impression of the resulting equations, we depicted the update equations for the energy and the T_1-/T_2-amplitudes for the frequently used CCSD procedure in Figure 3.5. Evaluation for the "doubles" amplitudes $f(t)_{IJ}^{AB}$ is an N^6 step if one supposes that the number of virtual orbitals in the chosen discretisation is proportional to N. Note the most expensive summation step (but not the only N^6 step) is given by the term $\frac{1}{2} \sum_{CD} \langle AB \| CD \rangle t_{IJ}^{CD}$ contributing to $f(t)_{IJ}^{AB}$ (see the second line). This step consists of a summation over V^2 terms for each of the N^2V^2 amplitudes, so that the evaluation of this contribution is a step of complexity N^2V^4, which usually consumes (due to the constants entering by $V = C \cdot N$) about $70 - 90\%$ of the overall computation time. Efficient factorization of the CC equations is another topic of its own, cf. e.g. the references in [56].

$$
\begin{aligned}
E(t) &= \langle \Psi_0, H\Psi_0\rangle + \sum_{IA} f_{IA} t_I^A + \frac{1}{4}\sum_{IJAB}\langle IJ\|AB\rangle t_{IJ}^{AB} + \frac{1}{2}\sum_{IJAB}\langle IJ\|AB\rangle t_I^A t_J^B,\\
f(t)_I^A &= f_{IA} + \sum_C f_{AC} t_I^C - \sum_K f_{KI} t_K^A + \sum_{KC}\langle KA\|CI\rangle t_C^K + \sum_{KC} f_{KC} t_{IK}^{AC}\\
&+ \frac{1}{2}\sum_{KCD}\langle KA\|CD\rangle t_{KI}^{CD} - \frac{1}{2}\sum_{KLC}\langle KL\|CI\rangle t_{KL}^{CA} - \sum_{KC} f_{KC} t_I^C t_K^A - \sum_{KLC}\langle KL\|CI\rangle t_K^C t_L^A\\
&+ \sum_{KCD}\langle KA\|CD\rangle t_K^C t_I^D - \sum_{KLCD}\langle KL\|CD\rangle t_K^C t_I^D t_L^A + \sum_{KLCD}\langle KL\|CD\rangle t_C^K t_{LI}^{DA}\\
&- \frac{1}{2}\sum_{KLCD}\langle KL\|CD\rangle t_{KI}^{CD} t_L^A - \frac{1}{2}\sum_{KLCD}\langle KL\|CD\rangle t_{KL}^{CA} t_I^D\\
f(t)_{IJ}^{AB} &= \langle IJ\|AB\rangle + \sum_C \left(f_{BC} t_{IJ}^{AC} - f_{AC} t_{IJ}^{BC}\right) - \sum_K \left(f_{KJ} t_{IK}^{AB} - f_{KI} t_{JK}^{AB}\right)\\
&+ \frac{1}{2}\sum_{KL}\langle KL\|IJ\rangle t_{KL}^{AB} + \frac{1}{2}\sum_{CD}\langle AB\|CD\rangle t_{IJ}^{CD} + P(IJ)P(AB)\sum_{KC}\langle KB\|CJ\rangle t_{IK}^{AC}\\
&+ P(IJ)\sum_C\langle AB\|CJ\rangle t_I^C - P(AB)\sum_K\langle KB\|IJ\rangle t_A^K\\
&+ \frac{1}{2}P(IJ)P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_{IK}^{AC} t_{LJ}^{DB} + \frac{1}{4}\sum_{KLCD}\langle KL\|CD\rangle t_{IJ}^{CD} t_{KL}^{AB}\\
&+ \frac{1}{2}P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_{IJ}^{AC} t_{KL}^{BD} - \frac{1}{2}P(IJ)\sum_{KLCD}\langle KL\|CD\rangle t_{IK}^{AB} t_{JL}^{CD}\\
&+ \frac{1}{2}P(AB)\sum_{KL}\langle KL\|IJ\rangle t_K^A t_L^B + \frac{1}{2}P(IJ)\sum_{CD}\langle AB\|CD\rangle t_I^C t_J^D\\
&- P(IJ)P(AB)\sum_{KC}\langle KB\|IC\rangle t_K^A t_J^C + P(AB)\sum_{KC} f_{KC} t_K^A t_{IJ}^{BC}\\
&+ P(IJ)\sum_{KC} f_{KC} t_I^C t_{JK}^{AB} - P(IJ)\sum_{KLC}\langle KL\|CI\rangle t_K^C t_{LJ}^{AB}\\
&+ P(AB)\sum_{KCD}\langle KA\|CD\rangle t_K^C t_{IJ}^{DB} + P(IJ)P(AB)\sum_{KCD}\langle AK\|DC\rangle t_I^D t_{JK}^{BC}\\
&+ P(IJ)P(AB)\sum_{KLC}\langle KL\|IC\rangle t_L^A t_{JK}^{BC} + \frac{1}{2}P(IJ)\sum_{KLC}\langle KL\|CJ\rangle t_I^C t_{KL}^{AB}\\
&- \frac{1}{2}P(AB)\sum_{KCD}\langle KB\|CD\rangle t_K^A t_{IJ}^{CD} + \frac{1}{2}P(IJ)P(AB)\sum_{KLC}\langle KB\|CD\rangle t_I^C t_K^A t_J^D\\
&+ \frac{1}{2}P(IJ)P(AB)\sum_{KLC}\langle KL\|CJ\rangle t_I^C t_K^A t_L^B - P(IJ)\sum_{KLCD}\langle KL\|CD\rangle t_K^C t_I^D t_{LJ}^{AB}\\
&- P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_K^C t_L^A t_{IJ}^{DB} - \frac{1}{4}P(IJ)\sum_{KLCD}\langle KL\|CD\rangle t_I^C t_J^D t_{KL}^{AB}\\
&+ \frac{1}{4}P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_K^A t_L^B t_{IJ}^{CD} + P(IJ)P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_I^C t_L^B t_{KJ}^{AD}\\
&+ \frac{1}{4}P(IJ)P(AB)\sum_{KLCD}\langle KL\|CD\rangle t_I^C t_K^A t_J^D t_L^B
\end{aligned}
$$

Figure 3.2: The CCSD equations for the CC energy $E(t)$ and for the T_1, T_2 amplitudes $f(t)_I^A$, $f(t)_{I,J}^{A,B}$, $(I, J \in \text{occ}, A, B \in \text{virt})$. In this, $P(IJ)f(I,J) := f(I,J) - f(J,I)$.

(iii) Newton's method for the CC function. To compute a root of the Coupled Cluster function (3.55), it is common practice to use an inexact Newton's method with the (lifted, shifted) Fock matrix as approximate Jacobian, or a related Jacobi-like scheme,[37] also cf. [56]. With the results of the previous sections, we now obtain a convergence result for the more general above setting, by which we close this present section.

Corollary 3.30. *(Convergence of a quasi-Newton method)*

Let $P : \mathbb{V} \to \mathbb{V}'$ be any linear bounded coercive linear mapping. If E^* is a simple eigenvalue and Ψ_0 is close enough to $\underline{\Psi}$, there is an $\alpha > 0$ such that the damped inexact Newton's method

$$x_{n+1} \quad = \quad x_n - \alpha P^{-1} f(x_n) \tag{3.56}$$

with starting value Ψ_0 converges to $\underline{\Psi}$. If $\|Df(t^*) - P\|$ is small enough, $\alpha = 1$ is a possible choice.

The proof is identical with that for the finite dimensional case, which can be derived e.g. from Theorem 8.2.2 in [60], so it is omitted.

□

[37] The Fock operator $F - \Lambda_0 I$ is related to the Jacobian of f, see [190] for the canonical orbital case and cf. the proof of Theorem 3.18, where it is shown that for a reference solution good enough, there holds $Df(t^*)g \approx (\langle \Psi_\mu (\hat{H} - E^* I) G \Psi_0 \rangle)_\mu$.

3.6 Concluding remarks

We have presented and analysed a well defined continuous Coupled Cluster method, resulting in a root problem for the Coupled Cluster function defined on a coefficient space $\mathbb{V}$ which reflects the original space $H^1 = \mathbb{H}^1_k$. Solutions of the root equation correspond to the exact eigenvectors of the weak eigenproblem, Problem 1.12. The CC equations for the discrete Hamiltonian $\hat{\mathbf{H}}$ from (1.82), normally used as starting point in quantum chemistry, can now be interpreted as a Galerkin discretisation of the continuous CC equations, and this ansatz has enabled us to formulate error estimates with respect to the continuous solution $\underline{\Psi}$, see Theorem 3.21 and Theorem 3.24. In particular, the error estimate from Theorem 3.24 provides a tool that might be used for error estimation with an appropriate refinement strategy. To this end, the quantities $\|t_d - t^*\|, \|z_d - z^*\|$ have to be estimated, and therefore, the discrete primal and dual problems (3.47) for (t_d, z_d) have to be solved, and an approxmation to the exact solution (t^*, z^*) of (3.46) has to be computed. This would for instance possible by using hierarchical basis sets as the VnZ-bases used in extrapolation schemes, or also by selecting subsets of a discretised set of amplitudes to estimate the effect of including e.g. only some of the T_2 amplitudes in a classical CCSD calculation. The practical design of such error estimators should be persued further in future work.

Our analysis also reflects the general weakness of the Coupled Cluster method if the spectral gap (3.39) is too small, or if multiple eigenvalues occur. In this case, multireference methods have to be utilized, and it would be desirable to use the theoretical framework developed here to attack this problem from the viewpoint of numerical analysis in the near future.

4 The DIIS acceleration method

Let us consider a (typically nonlinear) equation of the form

$$g(x^*) = 0. \tag{4.1}$$

In most iterative procedures, a residual-like correction term like $r_n = -g(x_n)$ or a preconditioned, damped or approximate variant of this is computed from the current iterate x_n, and the next update is then defined as $x_{n+1} := x_n + r_n$.[38] The DIIS (Direct Inversion in the Iterative Subspace) method, introduced by P. Pulay [170, 171], is an acceleration technique for solvers for nonlinear problems as (4.1), which exploits not only the information contained in x_n and r_n, but considers a number of previously computed iterates. Instead of letting $x_{n+1} := x_n + r_n$, DIIS lets $\tilde{x}_{n+1} := x_n + r_n$, and then computes in a supplementary step improved iterates

$$x_{n+1} = \sum_{i=\ell(n)}^{n} c_i \tilde{x}_{i+1} = \sum_{i=\ell(n)}^{n} c_i (x_i + r_i) \quad \text{with} \quad \sum_{i=\ell(n)}^{n} c_i = 1$$

by minimizing the least square functional

$$J_{DIIS}(y) \;:=\; \frac{1}{2} \Big\| \sum_{i=\ell(n)}^{n} c_i r_i \Big\|^2 \tag{4.2}$$

over the set of all coefficient vectors $(c_i)_{i=\ell(n)}^{n}$ for which $\sum_{i=\ell(n)}^{n} c_i = 1$. Usually, only a short history of previous iterates is considered, i.e. $n - \ell(n) + 1$ is a small number in the above.

DIIS was originally designed to accelerate the self consistent field iteration (SCF, cf. Sec. 2.3(ii)), but has been found to be quite useful in a much broader context to improve convergence for a variety of algorithms used in electronic structure calculations. We already mentioned that DIIS is used to enhance convergence of the direct minimization schemes introduced in Section 2, in particular in the context of DFT [83, 191] and is also utilized to speed up the iterative solution of the Coupled Cluster equations [103]. It does not only improve the iteration procedure significantly [130]; in the SCF iteration for DFT calculations, it even often turns non-convergent iterations into convergent ones. A variant of DIIS is used for simultaneous computation of eigenvalues and corresponding eigenvectors (RMM-DIIS, [126]) and has proven to be extremely efficient; when having to deal with the problem of *charge sloshing* that sometimes appears when DFT is applied to metallic systems, it seems to be superior to Broyden's method [126]. DIIS is not only popular in electronic structure calculation, but is also frequently used in molecular dynamics for

[38] Examples for such problems and corresponding iterations are for instance the minimization problems from Section 2, with (4.1) being the first-order optimality condition (2.14) for the functional $\mathcal{J}$ and the minimization algorithm Fig. 2.1 an allowing algorithm, or also the Coupled Cluster equations from Theorem 3.16, with the Quasi-Newton method (3.56) as a possible associated iterative method.

	basic iteration	with DIIS
DFT calculation for cinchonidine	43	22
CCSD calculation for N2, cc-pVTZ	21	12
CCSD calculation for LiH, cc-pVQZ	43	21

Figure 4.3: Iterations needed to converge some sample sample DFT/CCSD calculations with and without DIIS. (DFT calculation performed with bigDFT [30, 83], a part of the ABINIT package [1, 89, 90]), CC calculations performed with NWChem [41, 116].)

geometry optimization [66]. In these problems, comparisons with quasi-Newton methods, e.g. with BFGS, show that the two methods behave similarly, while BFGS seems to be slightly better when the problem under consideration is not well-conditioned [75]. Incorporation of ideas related to DIIS into the various physical applications of quantum chemistry has led to a further improvement and additional variants of DIIS, e.g. [57, 75, 102, 115, 217], without adding significant further costs. If the actual iterates are close to the solution, there are cases in which DIIS exhibits superlinear convergence.

Often, the basic algorithm $\tilde{x}_{n+1} = x_n + r(x_n)$ already produces a linearly convergent sequence of iterates, see e.g. Theorem 2.14 and Corollary 3.30 for examples and also cf. the remarks at the beginning of 4.1. In this case, DIIS normally approximately halves the number of iteration steps needed to reach a prescribed precision, see the sample calculations in Figure 4.3. This is the reason that DIIS is often termed a *convergence acceleration method*. Nevertheless, our analysis will show that there are cases where we do not have to assume the convergence of the basic algorithm $\tilde{x}_{n+1} := x_n + r(x_n)$ a priori to obtain convergence of DIIS, cf. Remark 4.8.

In the present Section 4, we will analyse the properties of DIIS from various viewpoints and show that it is connected to other well-known algorithms. Although similarities between DIIS and Newton-type methods are evident and have been used to improve the DIIS procedure [75, 102], the formal connection between them has to the author's knowledge not been worked out in all clarity in the literature. Therefore, we will start by showing in Section 4.2 that DIIS can be interpreted as a quasi-Newton method similar to Broyden's method [60] (Theorem 4.2) and set it in relation to other Broyden-like methods. As a model problem, the convergence behaviour of DIIS when applied to linear equations is investigated in Section 4.3. We establish a relation to the well-known GMRES scheme and use this relation to derive some (positive as well as negative) convergence estimates for DIIS applied to linear equations in Theorem 4.7. Section 4.4 then provides some convergence results for the nonlinear case. First of all, we prove in Theorem 4.10 that the DIIS procedure as given in Figure 4.4 is linearly convergent; in Theorem 4.12, we will use linear convergence and both the relation to Broyden-like methods and to the GMRES procedure to give a second, more refined convergence estimate.

In practice, "superlinear" convergence behaviour of DIIS is often observed in the sense that the ratio $\|r_{n+1}\|/\|r_n\|$ of successive residual norms decreases, and in the light of the analysis given here, we will discuss along the way the circumstances under which this behaviour can/cannot set in, see Sections 4.2(iii), 4.3(iv) and 4.4(i).

4.1 Notations and basic facts about DIIS

Throughout this section, we will denote by V a given Hilbert space with an inner product $\langle \cdot, \cdot \rangle$, and denote the induced norm by $\| \cdot \|$. The dimension of V will be denoted by D, where $D = \infty$ is admitted. We will be concerned with the root problem (4.1), where g is a function that maps the space V to itself. The more general case where W is another Hilbert space and a root of a function $g : V \to W$ has to be computed is included if we can assume that the Jacobian $J^* \in L(V, W)$ of g at x^* is invertible, and that we have a cheaply applicable preconditioner $P \in L(V, W)$ at hand that approximates J^* sufficiently well: Under these circumstances, P is also invertible, and instead of computing the roots of g, we turn our attention to the function $\tilde{g}(x) = P^{-1}g(x)$ that has the same roots as g, but maps $V \to V$. Also, the case where the basic iteration is damped or overrelaxated by a fixed parameter α is included by considering $\tilde{g}(x) = \alpha g(x)$.[39]

The DIIS procedure outlined above results in the algorithm from [170, 171], displayed in Figure 4.4.

Figure 4.4: The DIIS algorithm.

Initialization.

Function $g : V \to V$, starting value $x_0 \in V$, $n = 0$ given.

Loop over:

(1) Evaluate the residual $r_n = -g(x_n)$. Let $\tilde{x}_{n+1} := x_n + r_n$.

(2) Terminate if desired precision is reached, e.g. if $\|g(x_n)\| < \epsilon$.

(3) Choose a number of previous iterates $x_{\ell(n)}, \ldots, x_n$ to be considered during the DIIS procedure such that $g(x_{\ell(n)}), \ldots, g(x_n)$ are linearly independent.

(4) Compute
$$c_i = \operatorname{argmin}\Big\{ \Big\| \sum_{i=\ell(n)}^{n} c_i r_i \Big\|_2 \;\Big|\; \sum_{i=\ell(n)}^{n} c_i = 1 \Big\}. \tag{4.3}$$

(5) Let
$$x_{n+1} \;=\; \sum_{i=\ell(n)}^{n} c_i \tilde{x}_{i+1} \;=\; \sum_{i=\ell(n)}^{n} c_i x_i + \sum_{i=\ell(n)}^{n} c_i r_i, \tag{4.4}$$
and set $n \leftarrow n + 1$.

End of loop.

[39] Again, the preconditioned gradient steps of the direct minimization scheme (Fig. 2.1) for DFT-/CI-/eigenvalue computations and the ones of the Newton method (3.56) for CC serve as examples for a combination of both variants. In all these examples, V is some appropriate Sobolev space $H^1(\Omega)$ and $W = V'$.

The solution of the constraint minimization problem in step (4) is usually computed from by application of standard Lagrangian calculus to (4.3): this results in the linear system

$$\begin{pmatrix} \mathbf{B} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \lambda \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}, \tag{4.5}$$

with $\mathbf{B}$ determined by the matrix coefficients $b_{j,k} = \langle g(x_j), g(x_k) \rangle$, $\ell(n) \leq j, k \leq n$ and $\mathbf{1} = (1 \dots 1)$ a vector of length $n - \ell(n) + 1$, see [103] for an explicit derivation. In step (3) of the algorithm, $\ell(n)$, determining the number $n - \ell(n) + 1$ of previous iterates considered in the computation of x_{n+1}, will generally be fixed unless the system matrix $\mathbf{B}$ becomes ill-conditioned, in which case the number $\ell(n)$ is systematicly reduced, see e.g. [171] for details.

4.2 Equivalence of DIIS to a projected Broyden's method

(i) Rewriting DIIS as a Broyden's method. In the present section, we show that the DIIS algorithm from Figure 4.4 may be rewritten as a projected variant of a "reverse" Broyden's method,

$$x_{n+1} = x_n - H_n g(x_n),$$

wherein H_n is a secant approximation of the inverse of the Jacobian matrix of g at x_n, obtained from the previous iterates $x_{\ell(n)}, \dots, x_n$ and associated function evaluations $g(x_{\ell(n)}), \dots, g(x_n)$, and discuss the relation to other Broyden-type methods in parts (ii) and (iii). We need some preparations, taken care of next.

Definition 4.1. *(Spaces of differences)*

For a given sequence of iterates $x_0, x_1, \dots, x_n$ produced by DIIS, we define for $i = 0, 1, \dots, n-1$ the differences

$$s_i := x_{i+1} - x_i, \qquad y_i := g(x_{i+1}) - g(x_i) \tag{4.6}$$

as well for $n \geq 1$ as the spaces

$$K_n := \operatorname{span}\{s_i \mid i = \ell(n), \dots, n-1\}, \qquad Y_n := \operatorname{span}\{y_i \mid i = \ell(n), \dots, n-1\},$$

in particular, $K_n = Y_n := \emptyset$ if $\ell(n) = n$. We denote the orthogonal projector onto Y_n by Q_n. Finally, we define the projected differences

$$\hat{y}_0 := y_0; \qquad \hat{y}_n := y_n - \sum_{i=0}^{n-1} \frac{\hat{y}_i{}^T y_n}{\hat{y}_i{}^T \hat{y}_i} \, \hat{y}_i. \tag{4.7}$$

Theorem 4.2. *(Equivalence of DIIS and a projected Broyden's method)*
The compound iteration steps

$$x_n \to \tilde{x}_{n+1} \overset{DIIS}{\to} x_{n+1} \tag{4.8}$$

can equivalently be computed by a Broyden-like projected update formula

$$x_{n+1} \;=\; x_n - \big(C_n Q_n + (I - Q_n)\big) g(x_n) \;=:\; x_n - H_n g(x_n). \tag{4.9}$$

with the projector Q_n from Definition 4.1, and in which C_n is a secant approximation to the inverse of the Jacobian on the space of differences Y^n, fixed by

$$C_n \;=\; 0 \;\; on \;\; Y_n^{\perp}, \qquad C_n y_i = s_i \;\; for\ all \;\; i \in \{\ell(n), \ldots, n-1\}. \tag{4.10}$$

If $\ell(n) = 0$ in each step, so that the full history of iterates is considered, the DIIS inverse Jacobian H_n can be calculated from the Jacobian H_{n-1} by the rank-1 update formula

$$H_0 = I, \quad H_{n+1} \;=\; H_n \;+\; \frac{(s_n - H_n y_n)\hat{y}_n^T}{\hat{y}_n^T y_n}, \tag{4.11}$$

with the projected difference $\hat{y}_n$ defined in (4.7).

Before we approach the proof of Theorem 4.2 and then discuss the result in part (ii), we note at first that for arbitrary $n \in \mathbb{N}$, it is not hard to see that

$$\operatorname{span}\{g(x_{\ell(n)}), \ldots, g(x_n)\} \;=\; \operatorname{span}\{g(x_n), y_{\ell(n)}, \ldots, y_{n-1}\} \;=\; \operatorname{span}\{g(x_n), Y_n\}. \tag{4.12}$$

Therefore, the differences $y_{\ell(n)}, \ldots, y_{n-1}$ are linearly independent because $g(x_{\ell(n)}), \ldots, g(x_n)$ are by definition of the DIIS algorithm; in particular, the update formula (4.11) is well-defined.
We comprise some technical details needed for the proof of Theorem 4.2 in the next lemma. Note that (iii) shows the uniqueness of the solutions of the DIIS minimization task.

Lemma 4.3. *Let $n \in \mathbb{N}$ and a set of iterates $x_1, \ldots, x_n$ be fixed.*

(i) There holds for all $j \in \ell(n), \ldots, n-1$ that

$$K_n \;=\; \operatorname{span}\{x_i - x_j \mid j \neq i = \ell(n), \ldots, n-1\}, \tag{4.13}$$
$$Y_n \;=\; \operatorname{span}\{g(x_i) - g(x_j) \mid j \neq i = \ell(n), \ldots, n-1\}. \tag{4.14}$$

(ii) For any $\ell < n \in \mathbb{N}$, any set of vectors $v_\ell, \dots, v_n \in V$ and any set of coefficients $c_\ell, \dots, c_n \in \mathbb{R}$ for which $\sum_{i=\ell}^n c_i = 1$, we have

$$\sum_{i=\ell}^{n} c_i v_i \;=\; v_j + \sum_{\substack{i=\ell \\ i\neq j}}^{n} c_i (v_i - v_j) \tag{4.15}$$

for all $j \in \{\ell, \dots, n\}$, in particular;

$$x_j + K_n = \big\{ \sum_{i=\ell}^{n} c_i x_i \;\big|\; \sum_{i=\ell}^{n} c_i = 1 \big\}, \quad g(x_j) + Y_n = \big\{ \sum_{i=\ell}^{n} c_i g(x_i) \;\big|\; \sum_{i=\ell}^{n} c_i = 1 \big\} \tag{4.16}$$

for all such j.

(iii) There holds

$$\min \big\{ \;\big\| \sum_{i=\ell(n)}^{n} c_i g(x_i) \big\| \;\big|\; \sum_{i=\ell(n)}^{n} c_i = 1 \big\} \quad = \quad \|(I - Q_n) g(x_n)\|. \tag{4.17}$$

The minimizer $(c_i)_{i=\ell(n)}^n$ is unique and fulfils

$$\sum_{i=\ell(n)}^{n} c_i (g(x_i) - g(x_n)) \quad = \quad - Q_n g(x_n). \tag{4.18}$$

□

Proof. To prove (i), observe that for all $i \in \{\ell(n), \dots, n-1\}$,

$$s_i \;=\; x_{i+1} - x_i \;=\; x_{i+1} - x_j - (x_i - x_j) \;\in\; \operatorname{span}\{x_i - x_j \;|j \neq\; i = \ell(n), \dots, n-1\}$$

and that vice versa, $x_i - x_j = \sum_{k=i}^{j-1} s_k \in K^n$ if $i < j$, $x_i - x_j = -\sum_{k=j}^{i-1} s_k \in K^n$ if $i > j$, from which (4.13) follows. The proof for (4.14) is analogous. Equation (4.15) follows from the constraint condition $\sum_{i=\ell}^n c_i = 1$, yielding

$$\sum_{i=\ell}^{n} c_i v_i \;=\; v_j - (1 - c_j) v_j + \sum_{\substack{i=\ell \\ i\neq j}}^{n} c_i v_i \;=\; v_j + \sum_{\substack{i=\ell \\ i\neq j}}^{n} c_i (v_i - v_j)$$

for all $j \in \{\ell, \dots, k\}$. In particular, (4.16) follows from this together with (i), and implies

$$\inf \big\{ \;\big\| \sum_{i=\ell(n)}^{n} c_i g(x_i) \big\| \;\big|\; \sum_{i=\ell(n)}^{n} c_i = 1 \big\} \quad = \quad \inf\{ \;\|g(x_n) - y\| \;\; | \; y \in Y_n\},$$

from which (4.17), (4.18) can be concluded from the best approximation properties of Hilbert spaces. Finally, (ii) together with (4.12) and the linear independence of $g(x_{\ell(n)}), \dots, g(x_n)$ implies in particular that the vectors $g(x_n) - g(x_i), i = \ell(n), \dots, n-1$ are linearly independent, so that the minimizer $(c_i)_{i=\ell(n)}^n$ is unique as coefficient vector of the best approximation of $g(x_n)$ in Y_n.

Proof of Theorem 4.2. By linearity, there follows that $C_n(g(x_i) - g(x_n)) = x_i - x_n$ for $i = \ell(n), \dots, n-1$, cf. the proof of Lemma 4.3. Using the definition of the DIIS iterates and Lemma 4.3, we obtain

$$\begin{aligned}
x_{n+1} &= \sum_{i=\ell}^{n} c_i \tilde{x}_{i+1} = \sum_{i=\ell(n)}^{n} c_i x_i - \sum_{i=\ell(n)}^{n} c_i g(x_i) \\
&= x_n + \sum_{i=\ell(n)}^{n-1} c_i(x_i - x_n) - \Big(g(x_n) + \sum_{i=\ell(n)}^{n-1} c_i\big(g(x_i) - g(x_n)\big)\Big) \\
&= x_n + C_n\Big(\sum_{i=\ell(n)}^{n-1} c_i\big(g(x_i) - g(x_n)\big)\Big) - \Big(g(x_n) + \sum_{i=\ell(n)}^{n-1} c_i\big(g(x_i) - g(x_n)\big)\Big) \\
&= x_n - C_n Q_n g(x_n) - (I - Q_n) g(x_n) =: x_n - H_n g(x_n).
\end{aligned}$$

This proves (4.9) and (4.10). To show (4.11), we note first of all that for each $n \in \mathbb{N}_0$, H_n is fixed on Y_n by the condition $H_n y_i = s_i$ for all $i = \ell(n), \dots, n-1$, while on $Y_n^\perp$, $H_n = I$. We show by induction that the same holds for (4.11), which we denote by

$$\hat{H}_0 = I, \quad \hat{H}_{n+1} = \hat{H}_n + \frac{(s_n - H_n y_n)\hat{y}_n^T}{\hat{y}_n^T y_n}$$

for a moment. For $n = 0$, the assertion holds because $Y_n = \emptyset$ and $\hat{H}_0 = I$ by definition. For $n \in \mathbb{N}$, we have for all $y \in Y_n^\perp$ that

$$\hat{H}_n y = \hat{H}_{n-1} y + \frac{(s_{n-1} - \hat{H}_{n-1} y_{n-1})\hat{y}_{n-1}^T}{\hat{y}_{n-1}^T y_{n-1}} y = y$$

because $\hat{y}_{n-1} \in Y_n$, so using the induction hypothesis, $\hat{H}_n = I$ on $Y_n^\perp$. Moreover, for all $i = 0, \dots, n-2$,

$$\hat{H}_n y_i = \hat{H}_{n-1} y_i + \frac{(s_{n-1} - \hat{H}_{n-1} y_{n-1})\hat{y}_{n-1}^T}{\hat{y}_{n-1}^T y_{n-1}} y_i = s_i + 0,$$

by induction hypothesis and definition of $\hat{y}_{n-1}$. Finally, for y_{n-1},

$$\hat{H}_n y_{n-1} = \hat{H}_{n-1} y_{n-1} + \frac{(s_{n-1} - \hat{H}_{n-1} y_{n-1})\hat{y}_{n-1}^T}{\hat{y}_{n-1}^T y_{n-1}} y_{n-1} = s_{n-1},$$

completing the proof.

□

The next lemma that will be needed later in the Section 4.3 on linear problems, but also holds in the nonlinear case.

Lemma 4.4. *If for fixed $n \in \mathbb{N}$, $\ell(i) = 0$ for all $i = 1, \dots, n$, i.e. the full history of previous iterates has been used in every previous step of the DIIS procedure and in particular, $g(x_0), \dots, g(x_{n-1})$ are linearly independent, there holds*

$$K_n \quad = \quad \operatorname{span}\{g(x_0), \dots, g(x_{n-1})\}. \tag{4.19}$$

Proof. We prove (4.19) by induction on n. For $n = 1$, $g(x_0) = x_1 - x_0$, so the statement holds in this case. For arbitrary $n \in \mathbb{N}$, we exploit (4.12) again, so that to show the assertion for $n+1$, it suffices to show that $x_n - x_{n+1} \in \operatorname{span}\{g(x_n), Y_n\}$ and that $\dim K_{n+1} = n + 1$: Using Theorem 4.2, we have

$$x_{n+1} - x_n \quad = C_n Q_n g(x_n) + (I - Q_n) g(x_n) \tag{4.20}$$

and the first term on the right side is an element of $K_n \subseteq \operatorname{span}\{g(x_0), \dots, g(x_{n-1})\}$ by definition of C_n and induction hypothesis, while the second is in $\operatorname{span}\{g(x_n), Y_n\}$ by the definition of the projector Q_n. Because $g(x_0), \dots, g(x_{n-1})$ are linearly independent, the second component on the right hand side of (4.20) (orthogonal to Y_n) is nonzero, implying with (4.12) that $\dim K_{n+1} = n + 1$. This completes the proof. □

(ii) Relation to other Broyden-type methods. Theorem 4.2 shows that the DIIS procedure can be interpreted as a quasi-Newton method in which the Newton step, consisting in the (usually computationally too expensive) solution of a sequence of linear systems

$$J(x_n) s_n = -g(x_n) \tag{4.21}$$

with $J(x_n)$ denoting the Jacobian of g at x_n, is replaced by letting $s_n = -H_n g(x_n)$. Herein, H_n is a rank-$(n - \ell(n) - 1)$-update of the identity, approximating $J^{-1}(x_n)$ by exploiting the information about $J^{-1}(x_n)$ contained in the sequence of former iterates $x_{\ell(n)}, \dots, x_n$ and according function values $g(x_{\ell(n)}), \dots, g(x_n)$: For all $\ell(n), \dots, n-1$, the directional derivatives $J(x_n)s_n$ are approximated by mapping the corresponding finite differences y_n of function values to s_n, see (4.10). In pursuing the ansatz of using differences of formerly calculated quantities to approximate the Jacobian $J(x_n)$ (or its inverse), DIIS is thus similar to the various variants of Broyden's method (see e.g. [60, 163]), and we will discuss this relation a little deeper in the following. For this comparison, we suppose that $\ell(n) = 0$ for each n-th step of DIIS, so that the full history of iterates is considered in each step until DIIS terminates.

In Broyden's original method [35], starting in our setting with the initial approximate Jacobian $B_0 = I$, the approximate Jacobian B_{n+1} is a rank-1-update of B_n that fulfils the secant condition

$$B_{n+1} s_n = y_n \tag{4.22}$$

and has the additional property that the Frobenius norm[40] $\|B_{n+1} - B_n\|_F$ is minimal among all such possible updates B_{n+1}. The update is given by

$$B_0 = I, \qquad B_{n+1} = B_n + \frac{(y_n - B_n s_n) s_n^T}{s_n^T s_n}.$$

Although Broyden's method does not retain the original quadratic convergence of the (exact) Newton method (4.21), it is q-superlinearly convergent, meaning that the sequence of quotients

$$q_n := \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} \tag{4.23}$$

is not only bounded by a constant $c < 1$ as in the case of (q-)linear convergence, but converges to zero (see [60] for the classical case and [93] for extended results on the operator case).

The DIIS-quasi-Newton method (4.9) is a combination of two variants of Broyden's method: The first one is the *reverse Broyden's method* in which the *inverse* $J(x^*)^{-1}$ of the Jacobian is approximated directly by successive rank-1-updates H_{n+1} fulfilling $H_{n+1} y_n = s_n$ and having minimal deviation with respect to the Frobenius norm from H_n, resulting in[41]

$$H_0 = I, \qquad H_{n+1} = H_n + \frac{(s_n - H_n y_n) y_n^T}{y_n^T y_n}.$$

Although this method is also termed as "bad Broyden's method" due to its convergence behaviour in practice, that is inferior to the above "forward" technique, the proof for q-superlinear convergence of the forward method can be modified to show that the reverse Broyden's method also converges q-superlinearly [36].
The second method related to (4.9) is a modification of the "forward" Broyden method, the *Broyden's method with projected updates* [82], developed further in [147]. It consists in the ansatz that the secant condition (4.22) should not only be fulfilled for the latest secant s_n, but by demanding

$$B_{n+1} s_i = y_i \quad \text{for all} \quad 0 \le i \le n, \tag{4.24}$$

while in contrast, the approximations B_{n+1} computed in Broyden's method need not fulfil the condition. This results in the formula

$$B_0 = I, \qquad B_{n+1} = B_n + \frac{(y_n - B_n s_n) \hat{s}_n^T}{\hat{s}_n^T s_n}, \tag{4.25}$$

[40] The Frobenius norm is only defined in finite dimensional spaces V; in infinite dimensional spaces, the difference $B_{n+1} - B_n$ has to be a Hilbert-Schmidt operator for a meaningful extension of this concept. See [93] for an alternative, more global characterization of the Broyden update in infinite dimensional spaces V.

[41] This yields a method different from the "forward" Broyden method, for which B_n^{-1} can be computed as a (different) rank-1 update of B_{n-1}^{-1} by the Sherman-Woodbury-Morrison formula, see e.g. [60] for an introduction and a comparison of both methods.

in which $\hat{s}_n$ is the orthogonalization of s_n against all previous differences s_i. The projected method has the advantage that when applied to linear problems, the exact solution is computed in the $(D+1)$-th step [82], a property that might also have positive effect on problems that are "close to linear" in the sense that the second order terms in the Taylor expansion are relatively small.[42] Comparison of (4.10) and (4.25) now shows that DIIS (with full history) is the *reverse variant of the projected Broyden's method*, and we note that the reverse method (i.e. DIIS) is also introduced in [82], Algorithm II', but not analysed further due to its practical behaviour which - as in the non-projected case - seems to be inferior to the forward method, see [82] also for comments on numerical tests. This is in agreement with the outcome of [102], in which the forward projected method from [82] is re-introduced as an improvement of DIIS, termed the KAIN (Krylov Accelerated Inexact Newton) solver. The interested reader is also referred to [118] and the references given therein for more related Newton-type algorithms using Krylov spaces spanned by finite differences.

(iii) Superlinear convergence of DIIS? - Part I. We conjectured that as from the "good/forward Broyden's method" to the "bad/reverse Broyden's method", we might transfer theoretical results on q-superlinear convergence on the projected forward method (given in [82]) to the projected reverse variant, i.e. to DIIS. Unfortunately, the proof of q-superlinear convergence given in [82] is erroneous. Because it is closely related to one of the main difficulties in the theoretical analysis of DIIS, we go a little further into detail here.

We already noted that the main difference between the "classical" and the "projected" Broyden method is that the classical method does not have the property (4.24), and the important point to note is that for this reason, the classical method allows for an *infinite* series of approximate Jacobians B_n related to the former B_{n-1} by a rank-1-update formula. This property is used to show for the "classical Broyden's method" the crucial condition

$$\lim_{n\to\infty} \frac{\|(B_{n+1}-J)s_{n+1}\|}{\|s_{n+1}\|} \quad \to \quad 0, \tag{4.26}$$

which together with $q-$linear convergence of the algorithm is equivalent to $q-$superlinear convergence of the algorithm [60]. In contrast, note that if the differences $s_i, i = 0, \ldots, n$ grow linearly dependent, the projected update formula (4.25) is undefined because $\hat{s}_n = 0$; the same holds for the differences $y_i, i = 0, \ldots, n$ and the DIIS formula (4.11). In [82], this problem is resolved for the "projected Broyden" by restarting the algorithm every time this happens, so that the algorithm is reset at latest when the space dimension D is reached, and the analysis given in [82] relies heavily on the occurence of restarts to reduce the errors associated with the secant approximations on the spaces Y_n. On the other hand, the criterion (4.26) is proven in [82] by the implicit assumption that there

[42] It can also be shown that the classical Broyden's method computes the exact solution to a linear problem after $2D$ steps, see [81].

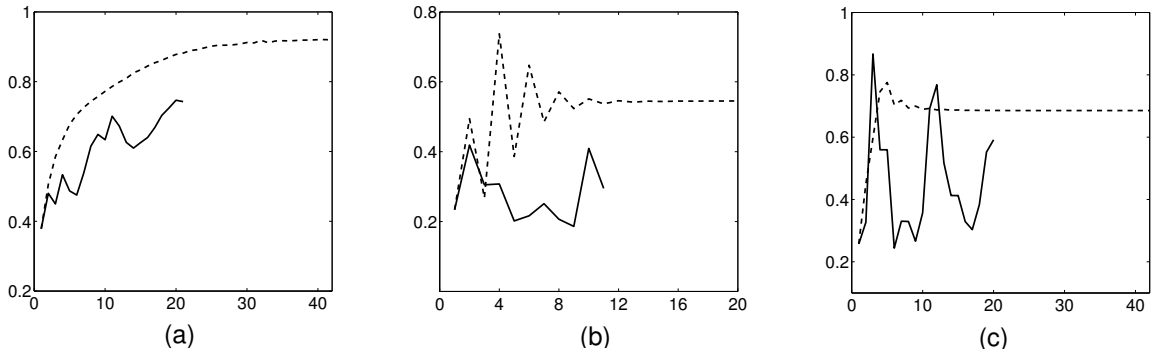

Figure 4.5: Ratio of the residuals $\|r(x_{n+1})\|/\|r(x_n)\|$ in the course of the iteration for the sample calculations displayed in Figure 4.3. (a) DFT for cinchonidine, (b) CCSD for N_2, (c) CCSD for LiH. Dashed line: basic iteration only, solid line: with DIIS acceleration.

is an *infinite* series of updates produced according to (4.25), relating for each $n \in \mathbb{N}$ the Jacobian B_{n+1} to B_n - which, under the just discussed circumstances with necessarily occurring restarts, is impossible.
Unfortunately, this flaw is not straight-forward to mend, and it is unclear whether q-superlinear convergence can be shown at all for the DIIS procedure and the above Broyden's method with projected updates without imposing further conditions e.g. on the Jacobian $J(x^*)$; also, in order to formulate results like q-superlinear convergence (as a limit process for $n \to \infty$), a suitable replacement/discarding strategy for former iterates that are "almost linearly dependent" will have to be formulated instead of just restarting the algorithm, which in practice results in a maximal cycle of D successive iterations.

We will take a little different approach here and investigate the transient (i.e. "short-term") convergence behaviour of DIIS by treating it as a nonlinear variant of the well-known GMRES procedure. Although sometimes in practical DIIS calculations, "superlinear" convergence behaviour can be observed in the sense that the ratio $\|r(x_{n+1})\|/\|r(x_n)\|$ of the residuals decays with increasing iteration number n, our general experience with DIIS is rather reflected exemplarily in Figure 4.5, where for the sample calculations from Figure 4.3, the above ratio has been plotted against the number n of iterations.

In our theoretical analysis in Section 4.4, we will find that the worst-case short-term convergence behaviour of DIIS essentially depends on balancing two opposing error terms associated with the number of previous iterates considered for DIIS, see Remark 4.13; cf. also part (iv) of the next section.

□

4.3 DIIS applied to linear problems

(i) Viewpoint and assumptions. As a model problem, we will now investigate the special case where DIIS is applied to a linear equation, i.e. for $A : V \to V$ linear and bounded, $b \in V$, an $x^* \in V$ is sought such that

$$g(x^*) = Ax^* - b \;=\; 0. \tag{4.27}$$

We use the negative gradient direction as update directions,

$$r(x_n) \;:=\; -\,g(x_n) \;=\; b - Ax_n. \tag{4.28}$$

By modifying g appropriately (see the remarks in Section 4.1), preconditioned or damped gradients used in the basic iteration scheme are also included; also, weak equations are covered. If we suppose that the iteration scheme is convergent, it is a well-known problem that convergence of this scheme can be extremely slow, especially if the condition number of A is large (see e.g. [96]). To overcome these problems, procedures like the well-known GMRES or cg solvers were developed, leading to accelerated convergence of the underlying scheme. We will now show that DIIS now inherits some of these properties from GMRES because the minimization problems solved and the subspaces used coincide.

We will assume that A is invertible, so that x^* is unique. Also, we will assume that in each step (2) of the DIIS algorithm displayed in Fig. 4.4, the full set of previous vectors $x_0, \ldots, x_n, g(x_0), \ldots, g(x_n)$ is used to minimize the least square functional (4.2). Note that this implies that the vectors $g(x_0), \ldots, g(x_n)$ have to be linearly independent, and we will see later that this indeed is the case unless $g(x_n) = 0$, in which case the algorithm terminates.

In the n-th step of a DIIS-accelerated gradient solver, the functional (4.2) is minimized over the space $x_n + K_n$. For the linear problem (4.27), this minimizer is by Lemma 4.3 and Lemma 4.4 given by

$$\bar{x} \;=\; \operatorname{argmin}_{\sum_{i=1}^n c_i = 1} \Big\{ \Big\| \sum_{i=0}^{n} c_i g(x_i) \Big\|^2 \Big\} \;=\; \operatorname{argmin}_{(c_i)_{i=0}^{n-1} \in \mathbb{R}^n} \Big\{ \Big\| A(x_0 + \sum_{i=0}^{n-1} c_i r_i) - b \Big\|^2 \Big\};$$

and using linearity once more, it is not hard to see that the next DIIS iterate is given by

$$x_{n+1} \;:=\; \sum_{i=0}^{n} c_i \tilde{x}_{i+1} \;:=\; \sum_{i=0}^{n} c_i x_i + \sum_{i=0}^{n} c_i g(x_i) \;=\; \bar{x} + r(\bar{x}). \tag{4.29}$$

For a comparison of DIIS with Krylov subspace methods, we now introduce for a given starting value $v_0 \in V$, $n \geq 1$ the well-known Krylov spaces

$$K_n(A, r(v_0)) \;:=\; \operatorname{span}\{A^i r(v_0) : i = 0, \ldots, n-1\}.. \tag{4.30}$$

We remind the reader that one point of view on Krylov subspace methods for linear systems is that they consist in iterating the two following steps:

(i) Minimize a given error functionals J over the space $v_0 + K_n(A, r(v_0))$ to obtain the next iterate $v_n \in v_0 + K_n(A, r(v_0))$.

(ii) Compute $A^n r(v_0)$ and construct the next Krylov space $K_n(A, r(v_0))$.

If A is symmetric, the well-known method of conjugate gradients ("cg") is an example for such a method, consisting in minimization of the functional

$$J_{\mathrm{cg}}(y) \;=\; \frac{1}{2}\langle A(y - x^*), y - x^*\rangle. \tag{4.31}$$

over the respective affine Krylov spaces $v_0 + K_n(A, r(v_0))$. With $\tilde{b} = b - Av_0$, and the minimizer written as $v_n = v_0 + \delta_n$, $\delta_n \in K_n(A, r(v_0))$, the first order condition for (4.31) is the Galerkin (orthogonality) condition $A\delta_n - \tilde{b} \perp K_n$, or more explicitly, $\langle A\delta_n - \tilde{b}, v\rangle = 0$ for all $v \in K_n$. Another example is given by the least-squares functional

$$J_{LS}(y) \;=\; \frac{1}{2}\langle A(y - x^*), A(y - x^*)\rangle \;=\; \frac{1}{2}\|Ax - b\|^2. \tag{4.32}$$

for which the first order condition for (4.32) is given by the Petrov-Galerkin condition

$$\langle A\delta_n - \tilde{b}, Av\rangle \;=\; 0 \quad \text{for all} \quad v \in K_n. \tag{4.33}$$

This is an oblique projection method [185] with

$$A\delta_n - \tilde{b} \;\perp\; AK_n, \tag{4.34}$$

i.e. the residual $A\delta_n - \tilde{b}$ of the optimal subspace solution v_n is A-orthogonal to K_n, or, in other words, the difference $v_n - x^*$ to the true solution x^* is A^2-conjugate to K_n. The Krylov method associated with (4.32) is the well-known GMRES-method [185], which for symmetric matrices results in the method of conjugate residuals ("cr", see e.g. [96]).
Let us note for later purposes that the Krylov spaces (4.30) allow for the alternative characterization

$$K_n(A, r(v_0)) \;=\; \text{span}\,\{r(v_0), \ldots, r(v_{n-1})\}, \tag{4.35}$$

see e.g. [96], Theorem 9.4.2 for a proof.

(ii) Connection between DIIS and GMRES. Comparison of the functionals (4.32) and (4.2) shows that the functionals used in GMRES and DIIS coincide. We will now clarify the relation between DIIS and GMRES, and thus also between GMRES and the projected Broyden's method from Theorem 4.2, further. Although Broyden-like secant methods have been proposed as an alternative to GMRES to solve large-scale linear equations (see [61] and references therein), the author is not aware of literature where the below connection between GMRES and the projected Broyden's method from Theorem 4.2 is made explicit.

Lemma 4.5. *If the starting values of a GMRES procedure and a DIIS procedure applied to the linear system (4.27) coincide, $x_0 = v_0$, there holds*

$$K_n(A, r(v_0)) = K_n \tag{4.36}$$

for any $n \in \mathbb{N}$. The GMRES procedure and the DIIS procedure, applied to linear problems, therefore solve the same minimization problem in each step (only using a different parametrization). The iterates x_n of the DIIS procedure and the iterates v_n of GMRES are related by

$$x_{n+1} = v_n - r(v_n). \tag{4.37}$$

There holds

$$\|r(v_{n+1})\|_2 \quad \leq \quad \|r(x_{n+1})\|_2 \quad \leq \quad \|I - A\|_2 \|r(v_n)\|_2. \tag{4.38}$$

□

In Figure 4.6, the result of Lemma 4.5 is displayed in a flow chart comparing GMRES and DIIS; the iterates of DIIS are denoted by x_n, those of GMRES by v_n.

Figure 4.6: The (linear) DIIS procedure vs. the GMRES algorithm

---

Initialization.

▷ Starting value $x_0 = v_0 \in V$, $n = 1$ given. Compute $r_0 = r(x_0)$, let $K_1 := \operatorname{span}\{d_0\}$.

▷ DIIS: Set $x_1 := x_0 + r(x_0)$. Compute $r_1 := r(x_1)$.
GMRES: Compute $r_1 := Ar_0$ from $r(v_0)$.

Loop over:

(1) Add r_n to K_n to obtain K_{n+1}. (The spaces K_n coincide for GMRES and DIIS, Lemma 4.5.)

(2) Calculate $\bar{x} \in x_0 + K_{n+1}$ which minimizes the residual over $x_0 + K_{n+1}$.

(3) DIIS: Let $x_{n+1} = \bar{x} - r(\bar{x})$. Compute $r_{n+1} = r(x_{n+1})$.
GMRES: Let $v_n = \bar{x}$. Compute $r_n := Ar(v_{n-1})$.

(4) Set $n \leftarrow n + 1$.

End of loop.

Proof. We use the representation (4.35) for the Krylov spaces, and the analogous one from Lemma 4.4 for the spaces used in DIIS, $K_n = \operatorname{span}\{r(x_0), \dots, r(x_{n-1})\}$. We proceed by induction. For $n = 1$, $K_1(A, r(v_0)) = \operatorname{span}\{r(v_0)\} = \operatorname{span}\{r(x_0)\} = K_1$ holds trivially, and $x_1 = v_0 - r(v_0)$ holds by definition of DIIS. Now let the assertion hold for fixed $n \in \mathbb{N}$. We then get that

$$K_{n+1} = \operatorname{span}\{K_n, r(x_n)\}, \qquad K_{n+1}(A, r(v_0)) = \operatorname{span}\{K_n(A, r(v_0)), r(v_n)\},$$

so that with the induction hypothesis, it suffices to show that $r(x_n) \in K_{n+1}(A, r(v_0))$ and that $r(v_n) \in K_{n+1}$. Using the induction hypothesis $x_n = v_{n-1} - r(v_{n-1})$, we have

$$r(x_n) \;=\; A(v_{n-1} + r(v_{n-1})) - b \;=\; r(v_{n-1}) + Ar(v_{n-1});$$

the first term to the right is in $K_n(A, r(v_0))$ according to (4.35), while the second is in $K_{n+1}(A, r(v_0))$ according to (4.30), so $r(x_n) \in K_{n+1}(A, r(v_0))$ follows. Vice versa,

$$Ar(v_{n-1}) = r(x_n) - r(v_{n-1}) \in K_{n+1}$$

because $r(x_n) \in K_{n+1}$ and $r(v_{n-1}) \in K_n(A, r(v_0)) = K_n$. Thus, $K_{n+1} = K_{n+1}(A, b)$, and because the functionals (4.32) and (4.2) also coincide, DIIS and GMRES both compute the same minimizer $\bar{x}$ on $x_0 + K_{n+1}$. While GMRES sets $v_{n+1} = \bar{x}$ by definition, we have $x_n = \bar{x} - g(\bar{x})$ in DIIS, see (4.30). This shows (4.37).

For the left inequality of (4.38), note that $x_{n+1} \in x_0 + K_n$, and that v_n minimizes the 2-norm of the residual over that space. The estimate on the right hand side follows directly from (4.37).

□

Lemma 4.5 shows that we can interpret GMRES as a variant of the DIIS/projected Broyden method for linear problems, exhibiting in the symmetric case the well-known advantages like the shortening of history [96]. While in the linear cases, the Krylov spaces (4.30) and the space (4.35) containing the current residuals coincide, this is not the case anymore in the case of nonlinear problems, and the residuals $g(x_n)$ then have to be evaluated explicitly, leading to the DIIS method. DIIS can thus be interpreted as a globalization of the least square ansatz of GMRES to the nonlinear case. Because in GMRES, only the former two iterates have to be respected to compute the residual minimizer over the whole Krylov space, it will be interesting to investigate how the omission of former iterates influences the convergence of DIIS applied to mildly nonlinear problems, and this is postponed to future work.
As a first corollary, Lemma 4.5 implies the following termination property of "linear DIIS".

Corollary 4.6. *In exact arithmetic, the DIIS procedure, applied to the iteration scheme $\tilde{x}_n = x_n - r(x_n)$ for the linear equation (4.27), terminates after $n \leq D$ steps with the exact solution $x_n = x^*$.*

Proof. Let us note at first that from Lemma 4.5, the vectors (4.19) building the spaces K_n become linearly dependent if and only if the vectors in (4.35) become linearly dependent. It is well-known [96, 185] that for GMRES, there holds $r(v_i) \perp_A K_i(A, b)$ for all $i \in \mathbb{N}$. In particular, the vectors in (4.35) become linearly dependent if and only if $r(v_i) = 0$ and $v_i = x^*$ is the solution of (4.27), and this will happen at latest when $i = D - 1$. For the corresponding DIIS iterate, there then holds $x_{i+1} = v_i + r(v_i) = x^*$ by (4.37), completing the proof.

□

(iii) Convergence of DIIS for linear problems. We now transfer well-known convergence properties of GMRES [141] to analyze the convergence behaviour of DIIS for the model problem of linear equations. Theorem 4.7 shows that as for GMRES, the worst-case convergence behavior of DIIS applied to normal matrices A is completely determined by the spectrum of A. In the nonnormal case however, the convergence behavior of the GMRES method may not be related to the eigenvalues in any simple way and understanding the convergence of GMRES in the general non-normal case still remains a largely open problem, and this property is thus also inherited by the DIIS procedure. The application of DIIS to nonnormal matrices A also allows for the counterexample (iii), which also has some implications for the discussion of superlinear convergence of DIIS. See the proof and below for more remarks.

Theorem 4.7. *(Convergence of DIIS applied to linear problems)*

(i) Let $\|I - A\| = \xi$. If A is symmetric positive definite, and

$$\gamma\|x\|^2 \quad \leq \quad \langle Ax, x\rangle \quad \leq \quad \Gamma\|x\|^2$$

holds for all $x \in V$, the residuals of DIIS obey the estimate

$$\|r(x_{n+1})\| \quad \leq \quad \xi \, \frac{2c^n}{1 + c^{2n}} \, \|r(x_0)\|, \tag{4.39}$$

in which c is given by $c = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1) < 1$, $\kappa := \Gamma/\gamma$.

(ii) If A is diagonalizable with $A = XDX^{-1}$, where D is a diagonal matrix containing the eigenvalues of A, and if the eigenvalues of A are contained in an ellipse with center c, focal distance d and semimajor axis a which excludes the origin, we let $\kappa(X) = \|X\|_2\|X^{-1}\|_2$ be the condition number of X and there holds

$$\|r(x_{n+1})\| \quad \leq \quad \xi \cdot \kappa(X) \, \frac{T_n(\frac{a}{d})}{T_n(\frac{c}{d})} \, \|r(x_0)\|, \tag{4.40}$$

with T_n denoting the n-th Chebyshev polynomial and ξ as in (i). In particular, if A is normal, the estimate (4.40) holds with $\kappa(X) = 1$.

(iii) *Suppose we are given a nonincreasing sequence of D positive numbers*

$$r_0 \;\geq\; \ldots \;\geq\; r_{D-1} \;>\; 0$$

and D complex numbers $\lambda_1, \ldots, \lambda_D$. Then there exists a matrix $A \in \mathbb{C}^{D\times D}$ having the eigenvalues $\lambda_1, \ldots, \lambda_D$, a starting value x_0 and a right hand side b such that DIIS, applied to the tuple (A, b, x_0), gives a sequence of iterates $x_0, x_1, \ldots, x_D = x^*$ for which

$$\|r(x_i)\| \;\geq\; r_i \quad \text{for all} \quad 0 \;\leq\; i \;\leq\; D-1.$$

Proof. Theorem 4.7 follows together with Lemma 4.5, Eq. (4.38) from the respective properties of the GMRES procedure: Under the assumptions made in (i), there holds

$$\|Av_n - b\| \quad\leq\quad \frac{2c^n}{1+c^{2n}}\,\|Av_0 - b\|, \tag{4.41}$$

for the iterates of GMRES, see e.g. [96], Theorem 9.5.6 for the proof; the results also globalize straightforwardly to the operator case. The analogous estimate for the case (ii) where A may only be diagonalizable is for instance proven in [185], Proposition 6.32 and Corollary 6.33. The counterexample (iii) is a reformulation of the central result of [11], where an analogous statement is proven for GMRES.

□

Remark 4.8. (DIIS as acceleration method for linear systems)
Theorem 4.7 gives an insight on how DIIS accelerates convergence in the linear case: While the basic iteration scheme $x_n \leftarrow x_{n-1} - r_{n-1}$, for instance a simple (maybe damped) gradient algorithm, may converge slow or not at all, DIIS optimizes the residual over the whole space K_n, and thus inherits the nice convergence behaviour of GMRES. In particular, for the finite history of length 2, this leads to a line search over the space $x_{n-1} + \alpha r(x_{n-1}), \alpha \in \mathbb{R}$, so that in this case, DIIS may turn non-convergent iterations into convergent ones as a consequence of the convergence of the Richardson iteration with properly chosen α. We note that this behaviour is also sometimes observed when DIIS is applied to nonlinear systems.

(iv) Superlinear convergence, part II: Conclusions from the linear case. When applied to finite dimensional systems, the DIIS method provides the exact solution after at most $n = D$ steps according to Lemma 4.6. The general notion of (super-)linear convergence (as a limit process for $n \to \infty$, see 4.2(ii)) is therefore not appropriate for examination of the convergence behaviour in this case. An alternative that is also of more practical interest is the examination of how fast the sequence of DIIS residual norms $\|r(x_n)\|$ decays in the course of a moderate number $n \ll D$ of iterations.
The DIIS scheme essentially reproduces the convergence behaviour of the GMRES scheme,

for which in many cases some kind of "superlinear" convergence behaviour can be observed in practice in the sense that the ratio $\|r(x_{n+1})\|/\|r(x_n)\|$ of the residuals decays in the course of the iteration [110], and some results on circumstances under which the GMRES algorithm exhibits in some sense superlinear convergence are available: In [204], it is shown that the decay of the residual norms can be related to how well the outer eigenvalues of A are approximated by the Ritz values of A on the trial subspaces K_n; to the authors' knowledge, there is no analysis available though under which circumstances this approximation property is given. Other approaches relate superlinear convergence behaviour to certain properties of the a priori information provided by the data A, b and x_0, see e.g. [23, 24] for corresponding results for the related [96] cg-method.
Neverless, Theorem 4.7 (iii) displays that such "superlinear convergence behaviour" cannot always be expected for DIIS/GMRES, also cf. e.g. the last numerical example in [204].

4.4 Convergence analysis for DIIS

In this final section, we will give two convergence results for DIIS applied to nonlinear problems. We saw that DIIS can be reformulated as a projected Broyden method, and we at first will follow the lines of proof that are generally pursued in this context, and therefore as a first step prove that DIIS is locally linearly convergent if the underlying iteration has this property.
Linear convergence is then usually used to prove sharper results like superlinear convergence, see e.g. [60]. For the DIIS/projected reverse Broyden scheme, though, the corresponding proofs do not extend straightforwardly, cf. the remarks at the end of Section 4.2; moreover, Theorem 4.7 (iii) shows that if superlinear convergence can be shown for DIIS at all, there are cases where the superlinear convergence behaviour sets in after $n > D$ steps, while in the context of quantum chemistry, D is usually much larger than the number of maximal iteration steps. We will therefore show instead in Theorem 4.12 that DIIS combines the favourable properties of Newton's method with those of a GMRES solver applied to solve the actual linearized Newton's equation, where additional errors only arise from the error made in the finite difference approximation of the Jacobian $J(x^*)$.

(i) Assumptions and statement of the main results. Our analysis will be based on the following assumptions. Additionally to those which are standard in the analysis of quasi-Newton methods, we specify a more precise condition for the linear independence of former differences $y_{\ell(n)}, \ldots, y_{n-1}$ than was stated in the DIIS algorithm in Fig. 4.4.

Assumptions and Notations 4.9. *We assume the function $g : V \to V$ be differentiable in an open convex set $E \subseteq V$, and that $g(x^*) = 0$ holds for some $x^* \in E$. Denoting for $A \in L(V)$ its operator norm by $\|A\|$, we further assume that for some $K \geq 0$,*

$$\|g'(x) - g'(x^*)\| \quad \leq \quad K \ \|x - x^*\| \tag{4.42}$$

holds for all $x \in E$, and that the Jacobian $J := g'(x^*)$ is nonsingular. We will denote

$$\gamma \;:=\; \|J^{-1}\| \;=\; \|g'(x^*)^{-1}\|. \tag{4.43}$$

We will also assume that

$$\|I - J^{-1}\| < \delta \tag{4.44}$$

is sufficiently small. If this is not the case, we can use the function $\tilde{g}(x) = P^{-1}g(x)$ instead, where P is an approximation of J, and the above condition is then replaced by the condition that g can be preconditioned sufficiently well such that $\|I - J^{-1}P\| < \delta$.

Finally, we will assume that for the sequence of former iterates $x_{\ell(n)}, \ldots, x_n$ considered in the step $n \to n+1$, the corresponding differences of function values fulfil

$$\|P_{j\neq i} y_i\| \quad \geq \quad \frac{\|y_i\|}{\tau} \quad \text{for all} \quad i = \ell(n), \ldots, n-1 \tag{4.45}$$

for some $\tau > 1$, where $P_{j\neq i}$ denotes the projector on

$$Y_{n,j\neq i} \;=\; \mathrm{span}\{y_j | i = \ell(n), \ldots, n-1, j \neq i\}.$$

□

Note that results analogous to the ones below also hold if the Lipschitz condition (4.42) is replaced by a more general Hölder condition as used e.g. in [163, 60]. Because the functions used in quantum chemistry are usually locally Lipschitz continuous (see Sec. 2.2, Sec. 3.3), we refrained from this generalization here.

The first convergence result we prove is that the DIIS method is ($q-$)linearly convergent for sufficiently good starting values. The according result is stated in the next theorem.

Theorem 4.10. *(Linear convergence of DIIS)*

Let $x_0, x_1, \ldots,$ be a sequence of iterates produced by DIIS update scheme from Fig. 4.4 – or equivalently, computed from (4.9) – , where in each step n, the number of former iterates $y_{\ell(n)}, \ldots, y_n$ used to build the subspace K_n is chosen such that the linear independence condition (4.45) is fulfilled.

Then, the sequence $x_0, x_1, \ldots,$ is locally linearly ($q-$)convergent for any $0 < q < 1/(2\tau)$, i.e. there are constants $\delta = \delta(q), \epsilon = \epsilon(q) > 0$ such that if $\|I - J^{-1}\| \leq \delta$, $\|x_0 - x^*\| \leq \epsilon$, we have $x_n \in E$ and there holds

$$\|x_{n+1} - x^*\| \quad \leq \quad q \cdot \|x_n - x^*\| \tag{4.46}$$

for all $n \in \mathbb{N}$.

The proof for Theorem 4.10 will be given in part (ii) of the present section. Our second convergence result, to be formulated in Theorem 4.12 and proven in part (iii) of this section, shows that DIIS can be interpreted as a quasi-Newton method, in which the Newton equation (4.47) is solved approximately by a GMRES/DIIS step for the linear system, and in which the Jacobian J (resp. $J(x_n) = g'(x_n)$) is approximated by finite differences, see also the remarks below. We introduce the necessary notation in the next definition.

Definition 4.11. *Let $n \in \mathbb{N}$ be fixed and let us denote by z^* the exact solution of the linear equation*

$$Jz^* \quad = \quad Jx_n - g(x_n) \quad =: \; b_n \tag{4.47}$$

By z_i, $\ell(n) \leq i \leq n+1$, we denote the iterates of a DIIS procedure applied to the linear equation (4.47) with starting value $z_{\ell(n)} := x_{\ell(n)}$. Thus,

$$z_{i+1} = z_i - G_i r(z_i),$$

in what $r(z_i) = Jz_i - b_n$ is the residual associated with the linear equation (4.47), and G_i is the DIIS inverse Jacobian, fulfilling

$$G_i\big(r(z_i) - r(z_{i+1})\big) = z_i - z_{i+1}$$

for all $\ell(n) \leq i \leq n$, see Theorem (4.2). We define the associated residual reduction factors,

$$d_{i-\ell(n)} \quad := \quad \frac{\|r(z_i)\|}{\|r(z_{\ell(n)})\|}.$$

In the case that $r(z_i) = 0$ for some $i = \ell(n), \ldots, \leq n+1$, we define $z_{i+j} := z_i$ for all $j \in \mathbb{N}$.

□

We can now formulate the announced second convergence estimate for DIIS under a little more restrictive assumptions, see also (ii) in the following remark.

Theorem 4.12. *(A refined convergence estimate for DIIS)*

Let the assumptions of Theorem 4.10 hold. Then there are $\delta = \delta(q), \epsilon = \epsilon(q) > 0$ such that if $\|I - J^{-1}\| \leq \delta$ and $\|x_0 - x^*\| \leq \epsilon$, and if $\ell(j) = \ell(n)$ for all $\ell(n) \leq j \leq n$, the "residual error" $\|g(x_{n+1})\|$ can be estimated by

$$\|g(x_{n+1})\| \quad \leq \quad c_1\|g(x_n)\|^2 \quad + \quad c_2 \cdot \; d_{n-\ell(n)} \; \|g(x_{\ell(n)})\| \quad + \quad c_3\|g(x_{\ell(n)})\|^2, \tag{4.48}$$

for all $n \in \mathbb{N}$, where $d_{n-\ell(n)}$ is the convergence factor obtained in the $(n-\ell(n))$-th step of the DIIS solution of the linear auxiliary problem from Definition 4.11.

Remark 4.13. (Notes on Theorem 4.12)

(i) In view of the idea and proof of Theorem 4.12, the three error components in estimate (4.48) have straight-forward interpretations:

- The first term represents the modeling (linearization) error of (the exact) Newton's method, where the correction equation (4.47)[43] is solved exactly, leading to quadratic convergence the well-known quadratic error term.
- The second term represents the error made in solving (4.47) approximately by a GMRES/DIIS step on the actual subspace $x_n + K_n$, thus incorporating the convergence rate of the DIIS/GMRES from Theorem 4.7.
- The third error term, that can grow large if many older iterates are regarded, is a worst-case estimate for the error made in the finite difference approximation of J^* resp. $J(x_n)$.

(ii) We conjecture that the latter error term can be bounded by $\|f(x_{\ell(n)})\| \cdot \|f(x_n)\|$, so that the result given here is presumably not optimal, but we were not able to show this so far. We also note that the restrictive assumption that $\ell(j) = \ell(n)$ for all $\ell(n) \leq j \leq n$ (meaning that in the DIIS procedure, $K_{\ell(n)} = \emptyset$, and that the used Krylov spaces K_j are constantly increased without discarding iterates; in particular, (4.45) has to be fulfilled in each step) could not be abolished without the error term $\|g(x_{\ell(n)})\|$ in the third term in (4.48) having to be replaced by the less favourable term $\|g(x_{\ell(\ell(n))})\|$.

(iii) We note that the second and third error term in (4.48) are opposing perturbations of the quadratic convergence given by the first term: The error term associated with the DIIS procedure for the linear problem (4.47) is reduced with an increasing number of former iterates, according to the well-known theory for the associated GMRES procedure, and thus gives better bounds the longer the history is chosen if the convergence of the GMRES procedure is favourable, e.g. superlinear. On the contrary, the error bound for the finite difference approximation gets worse the more former iterates are taken up in the procedure.
In order to obtain the best bounds for convergence rates for the DIIS procedure, the two error terms thus have to be balanced out, and in agreement with this, practical experience with GMRES seems to indicate that the number of iterates has to be kept moderate in order to keep the procedure efficient, especially if the iterates become "almost linearly dependent", i.e. if the constant τ gets large, see [115, 171]. Estimate (4.48) shows that such an inefficiency can solely be due to the effects of nonlinearity, contained in the third error term, so that in principle, if g is "rather

[43] Or alternatively, where the "real" Newton equation $J(x_n)(x_{n+1} - x_n) = F(x_n)$ is solved. Eq. (4.47) was chosen here for convenience, but it is not hard to see that replacing J by $J(x_n)$ only adds anther quadratic error term.

nonlinear" in the sense that the constant K in (4.42) is large, it is advisable to discard old iterates more often.

(iv) For linear problems, the first and last error terms in (4.48) are zero. By a continuity argument, we can heuristically conclude that if in contrast to the situation discussed in (iii), the nonlinearity, i.e. the constant K in (4.42), is small, the convergence of the DIIS is mainly governed by that of the associated DIIS/GMRES procedure for this problem. Note that in the context of electronic structure calculations, similar assumptions entered into our convergence analysis for CC and DFT, and they seem to be in good agreement with practice.
In particular, if the Jacobian is symmetric, for instance if (4.1) is the first order condition of a minimization problem as in DFT, the worst-case convergence behaviour of the DIIS procedure is mainly determined by the spectral properties of J, while for nonsymmetric Jacobians, properties of the right hand side etc. play a role, cf. Section 4.3.

(v) In particular, "superlinear convergence" of the algorithm can be expected if the DIIS/GMRES procedure for the underlying linear problem has this property already for a small number of steps, so that the third error term provoked by the nonlinearity of g and the associated finite difference approximation of J can be kept sufficiently small by discarding old iterates.

(ii) Proof of Theorem 4.10. In the present part of this section, we give the proof for the linear convergence of DIIS as asserted in Theorem 4.10. Although we proceed similarly to the analysis from [82] for the "forward" projected Broyden scheme, it should be noted that the bounds given there are improved significantly: The error terms in [82] are in the end bounded by $\text{const}\cdot(2\tau)^N$, where N is the dimension of the space, $\tau > 1$, and the neighbourhood $U(x^*, \epsilon)$ of the root x^* on which the procedure can be shown to be linearly convergent is determined by $\epsilon < (2\tau)^{-N}$. In the context of electronic structure calculations, where $N \approx 10^5 - 10^6$, this estimate is unsatisfactory, and we will show that it is possible to bound the according error terms without dependence on the dimension of the space.

The proof is preceded by some definitions, a remark collecting some general estimates, and two preparatory lemmas. Note that the recursion formula for calculation of H^n in reverse order (in contrast to the rank-1-update formula (4.11)) as well as the definition of the iterates $\bar{y}_i$ orthogonalized in reverse order (in contrast to (4.7)) have no practical meaning, but are merely used for theoretical purposes: In the investigation of the convergence behaviour of DIIS, it will help to show linear convergence for any sequence of DIIS inverse Jacobians, independent of the number of former differences y_k used in each step. Thus, we will not implicitly have to assume the occurrance of restarts as in [82], but can in every step choose an arbitrary number of former iterates fulfilling the linear independence condition (4.45).

Definition/Lemma 4.14. *For fixed $n \in \mathbb{N}$ and $Y_n := \operatorname{span}\{y_{\ell(n)}, \dots, y_{n-1}\}$, we introduce an orthogonal basis $\bar{y}_{\ell(n)}, \dots \bar{y}_{n-1}$ by orthonormalizing the basis of Y_n in descending order with the Gram-Schmidt procedure, i.e. by letting $\bar{y}_{n-1} = y_{n-1}$, and for $\ell(n) \le i \le n-2$,*

$$\bar{y}_i \quad = \quad y_i \quad - \quad \sum_{j=i+1}^{n-1} \frac{\bar{y_j}^T y_i}{\bar{y_j}^T \bar{y_j}} \bar{y}_j \quad =: \quad (I - Q_{i+1}^n) y_i.$$

Further, we define (again, in descending order)

$$H_n^n = I, \quad H_i^n := H_{i+1}^n + \frac{(s_i - H_{i+1}^n y_i)\bar{y}_i^T}{\bar{y}_i^T y_i} \quad \text{for} \quad \ell(n) \le i \le n-1. \tag{4.49}$$

Then, for H_n from (4.9), there holds

$$H_n = H_{\ell(n)}^n = I + \sum_{i=\ell}^{n-1} \frac{(s_i - H_{i+1}^n y_i)\bar{y}_i^T}{\bar{y}_i^T y_i}. \tag{4.50}$$

Moreover, we have

$$H_i^n y = H_j^n y \quad \text{for all} \quad y \in (\operatorname{span}\{y_i, \dots, y_{j-1}\})^\perp, \quad \ell(n) \le i < j \le n, \tag{4.51}$$

and with the quantities

$$\bar{s}_{n-1} := s_{n-1}, \qquad \bar{s}_i := s_i - H_{i+1}^n Q_{i+1}^n y_i \quad \text{for} \quad \ell(n) \le i \le n-2,$$

formula (4.49) can be rewritten as

$$H_n^n = I, \qquad H_i^n := H_{i+1}^n + \frac{(\bar{s}_i - H_{i+1}^n \bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T \bar{y}_i} \quad \text{for} \quad \ell(n) \le i \le n-1. \tag{4.52}$$

□

The proof is quite straightforward and very similar to the proof of (4.11) and of the analogous result in [82], so it is omitted.

Before we continue, we remark the following results well-known in analysis of quasi-Newton methods, see e.g. [60, 163] for the proofs of the finite-dimensional case, which also transfer directly to the infinite dimensional case in the form given here.

Remark 4.15. *From the assumptions stated in 4.9, we get that for all $u, v \in E$,*

$$\begin{aligned} \|g(v) - g(u) - J(u-v)\| &\le K\|v-u\| \max\{\|u-x^*\|, \|v-x^*\|\} && (4.53)\\ &\le 2K \left(\max\{\|u-x^*\|,\ \|v-x^*\|\}\right)^2; && (4.54) \end{aligned}$$

in particular, there holds for all $h \in V$ for which $x^* + h \in E$ that

$$\|g(x^*+h) - g(x^*) - Jh\| \quad \le \quad \frac{\gamma}{2}\ \|h\|^2. \tag{4.55}$$

Moreover, on a neighbourhood $U_\kappa(x^*)$, $\kappa > 0$, there holds for some $\rho > 0$ that

$$\frac{1}{\rho}\ \|v-u\| \quad \le \quad \|g(v) - g(u)\| \quad \le \quad \rho\ \|v-u\|. \tag{4.56}$$

The next supplementary result is a technical lemma which is an analogue (with improved constants) of Lemma 4.3. from [82].

Lemma 4.16. *Fix $n \in \mathbb{N}$. Let the assumptions from 4.9 hold, and define for $i \leq j \in \mathbb{N}$*

$$m_i^j \quad := \quad \max\{\|x_i - x^*\|,\ \|x_{i+1} - x^*\|, \ldots,\ \|x_j - x^*\|\},$$

and $c := (1+\delta)K\rho$. For the quantities $\bar{s}_i, \bar{y}_i$ from Definition 4.14, $\ell(n) \leq i \leq n-1$, there holds

$$\|\bar{s}_i - J^{-1}\bar{y}_i\| \quad \leq \quad c\Big(\sum_{j=i}^{n-1} m_j^{j+1}(2\tau)^{j-i}\Big)\|y_i\|. \tag{4.57}$$

with τ defined in (4.45).

Proof. We proceed by descending induction, starting from $i = n-1$. In this case, $\bar{s}_{n-1} = s_{n-1}$ and $\bar{y}_{n-1} = y_{n-1}$, so that the estimate (4.53) gives

$$\begin{aligned}\|\bar{s}_{n-1} - J^{-1}\bar{y}_{n-1}\| &\leq \|J^{-1}\|\ \ \|g(x_{n-1}) - g(x_n) - J(x_{n-1} - x_n)\| \\ &\leq (1+\delta)K\|s_{n-1}\|m_{n-1}^n \\ &\leq (1+\delta)K\rho\|y_{n-1}\|m_{n-1}^n.\end{aligned}$$

For $\ell(n) \leq i < n-1$, we get by definition of $\bar{s}_i, \bar{y}_i$ that

$$\begin{aligned}\|\bar{s}_i - J^{-1}\bar{y}_i\| &\leq \|s_i - J^{-1}y_i\| + \|H_{i+1}^n Q_{i+1}^n y_i - J^{-1}Q_{i+1}^n y_i\| \\ &\leq (1+\delta)K\|s_i\|m_i^{i+1} + \sum_{j=i+1}^{n-1}\|(H_{i+1}^n - J^{-1})\bar{y}_j\|\ |\frac{\bar{y}_j^T y_i}{\bar{y}_j^T\bar{y}_j}|, \\ &\leq c\|y_i\|m_i^{i+1} + \sum_{j=i+1}^{n-1}\|(H_{i+1}^n - J^{-1})\bar{y}_j\|\ \frac{\|y_i\|}{\|\bar{y}_j\|},\end{aligned}$$

where (4.53) was used again to estimate the first term, while the second is derived from the definition of the projector Q_{i+1}^n. Inserting for $\|(H_{i+1}^n - J^{-1})\bar{y}_j\| = \|\bar{s}_j - J^{-1}\bar{y}_j\|, j > i$, the induction hypothesis (4.57) and then using $\|y_j\|/\|\bar{y}_j\| \leq \tau$ yields

$$\begin{aligned}\|\bar{s}_i - J^{-1}\bar{y}_i\| &\leq c\|y_i\|m_i^{i+1} + c\sum_{j=i+1}^{n-1}\big(\sum_{k=j}^{n-1} m_k^{k+1}(2\tau)^{k-j}\|y_j\|\big)\frac{\|y_i\|}{\|\bar{y}_j\|} \\ &\leq c\|y_i\|\Big(m_i^{i+1} + \tau\sum_{j=i+1}^{n-1}\big(\sum_{k=j}^{n-1} m_k^{k+1}(2\tau)^{k-j}\big)\Big) \\ &= c\|y_i\|\Big(m_i^{i+1} + \tau\sum_{k=i+1}^{n-1} m_k^{k+1}\big(\sum_{j=i+1}^{k}(2\tau)^{k-j}\big)\Big).\end{aligned}$$

Using

$$\sum_{j=i+1}^{k}(2\tau)^{k-j} \quad \leq \quad \tau^{k-(i+1)}\sum_{j=i+1}^{k} 2^{k-j} \quad \leq \quad 2^{k-i}\tau^{k-(i+1)},$$

we then obtain

$$\begin{aligned}\|\bar{s}_i - J^{-1}\bar{y}_i\| &\leq c\|y_i\|\big(m_i^{i+1} + \sum_{k=i+1}^{n-1} m_k^{k+1}(2\tau)^{k-i}\big)\\ &\leq c\|y_i\|\big(\sum_{k=i}^{n-1} m_k^{k+1}(2\tau)^{k-i}\big).\end{aligned}$$

□

Lemma 4.17. *There holds*

$$H_n - J^{-1} = (I - J^{-1})(I - Q_n) + \sum_{i=0}^{n-1} \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}. \tag{4.58}$$

Let $q < 1/(2\tau)$ and suppose, for $0 \leq i \leq n-1$, $\|x_i - x^*\| \leq q^i\epsilon$. Then

$$\|\sum_{i=0}^{n-1} \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}\| \leq \alpha\epsilon, \qquad \|H_n - J^{-1}\| \leq \delta + \alpha\epsilon, \tag{4.59}$$

where $\alpha = c\tau(1-2\tau q)^{-1}(1-q^{-1})$ and $\|I - J^{-1}\| \leq \delta$.

Proof. Let us fix $n \in \mathbb{N}$. We use the representation from Definition/Lemma 4.14 and prove the estimate by descending induction on the matrices $H_n^n, H_{n-1}^n \ldots, H_{\ell(n)}^n = H_n$. For $i = n$, $H_n^n - J^{-1} = I - J^{-1}$, so the assertion is trivially true. For $\ell(n) \leq i \leq n-1$,

$$\begin{aligned}H_i^n - J^{-1} &= H_{i+1}^n - J^{-1} + \frac{(\bar{s}_i - H_{i+1}^n\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}\\ &= H_{i+1}^n - J^{-1} + \frac{(\bar{s}_i - J^{-1}\bar{y}_i + J^{-1}\bar{y}_i - H_{i+1}^n\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}\\ &= (H_{i+1}^n - J^{-1})\Big(I - \frac{\bar{y}_i\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}\Big) + \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i}.\end{aligned}$$

Thus, by induction and orthogonality of the vectors $\bar{y}_i$,

$$H_n - J^{-1} = H_{\ell(n)}^n - J^{-1} = (I - J^{-1})(I - Q_n) + \sum_{i=0}^{n-1} \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T\bar{y}_i},$$

showing the first claim (4.58). As per the second, we estimate this by (4.57) and use $m_j^{j+1} \leq q^j\epsilon$,

$$\begin{aligned}\|H^n - J^{-1}\| &\leq \|I - J^{-1}\| + \sum_{i=0}^{n-1} \frac{\|\bar{s}_i - J^{-1}\bar{y}_i\|}{\|\bar{y}_i\|} \leq \delta + c\tau\sum_{i=0}^{n-1}\Big(\sum_{j=i}^{n-1} m_j^{j+1}(2\tau)^{j-i}\Big)\\ &\leq \delta + c\tau\sum_{i=0}^{n-1}\Big(\sum_{j=i}^{n-1} q^j\epsilon(2\tau)^{j-i}\Big) \leq \delta + c\tau\sum_{i=0}^{n-1} q^i\epsilon\Big(\sum_{j=0}^{n-i-1}(2\tau q)^j\Big)\\ &\leq \delta + c\tau\epsilon(1-2\tau q)^{-1}(1-q^{-1}).\end{aligned}$$

□

We can now complete the proof for linear convergence with the help of the estimate (4.59).

Proof of Theorem 4.10. For given $0 < q < 1/(2\tau)$, we choose $\delta = \delta(q), \epsilon = \epsilon(q) > 0$ such that

$$(1+\delta)K\epsilon + \rho(\delta + \alpha\epsilon) \quad \leq \quad q, \tag{4.60}$$

with α given in Lemma 4.17, and in such a way that the open ball $U_\epsilon(x^*) \cap U_\kappa(x^*)$ of radius $\min\{\epsilon, \kappa\}$ lies in E. Note that the second condition implies $x_0 \in E$. We now show inductively that $\|x_{n+1} - x^*\| \leq r \cdot \|x_n - x^*\|$ and $x_{n+1} \in E$ for $n \in \mathbb{N}$. There holds

$$\begin{aligned}
\|x_{n+1} - x^*\| &= \|x_n - H_n g(x_n) - x^*\| \\
&= \|x_n - x^* - J^{-1}(g(x_n) - g(x^*))\| + \|(H_n - J^{-1})(g(x_n) - g(x^*))\| \\
&= \|J^{-1}\| \, \|g(x^*) - g(x_n) - J(x_n - x^*)\| + \|H_n - J^{-1}\| \, \|g(x_n) - g(x^*)\| \\
&\leq (1+\delta)K\|x^* - x_n\|^2 + \rho\|H_n - J^{-1}\| \, \|x_n - x^*\|, \qquad (4.61)
\end{aligned}$$

where we have used $\|J^{-1}\| \leq 1 + \delta$ and (4.55) to estimate the first term in the last line, and (4.56) for the second term. For the case that $n = 0$, this gives

$$\|x_1 - x^*\| \quad \leq \quad \big((1+\delta)K\epsilon + \delta\rho\big)\|x_0 - x^*\| \quad \leq \quad q \cdot \|x_0 - x^*\|$$

by the choice of δ, ϵ, in particular, this implies $x_1 \in E$. For arbitrary $n \geq 1$, we can inductively suppose that for $0 \leq i \leq n-1$, $\|x_i - x^*\| \leq q^i\epsilon$ holds, and therefore, (4.59) is valid. It follows

$$\|x_{n+1} - x^*\| \quad \leq \quad (1+\delta)K\|x^* - x_n\| \; + \; \rho(\delta + \alpha\epsilon)\|x_n - x^*\|.$$

Because $\|x_n - x^*\| \leq \quad q^n\epsilon$ by induction hypothesis, it follows that

$$\|x_{n+1} - x^*\| \quad \leq \quad \big((1+\delta)Kq^n\epsilon + \rho(\delta + \alpha\epsilon)\big)\|x_n - x^*\| \quad \leq \quad q \cdot \|x_n - x^*\|$$

by the choice (4.60) of δ, ϵ, again also implying $x_{n+1} \in E$ and completing the proof.

□

(iii) Proof of Theorem 4.12. Before we approach the proof, we again prove two auxiliary lemmas: Some estimates are provided in Lemma 4.18, and the lengthy proof of another estimate needed in the proof of Theorem 4.12 is outsourced to the preceding Lemma 4.19.

Lemma 4.18. *(Useful estimates)*

(i) For $r(x) = Jx - b_n$ (cf. (4.47)), there holds for all $x \in E$ that

$$\|r(x) - g(x)\| \quad \leq \quad 2K\big(\max\{\|x_n - x^*\|, \|x - x^*\|\}\big)^2. \tag{4.62}$$

(ii) For $x_n \in E$, the solution z^* of the helping equation (4.47) fulfils

$$\|z^* - x^*\| \quad \leq \quad \frac{\gamma^2}{2}\|x_n - x^*\|^2. \tag{4.63}$$

(iii) If for some $i \in \{\ell(n), \dots, n\}$ $x_i, z_i \in E$ and $\|x_i - z_i\| \leq c \cdot \|g(x_{\ell(n)})\|^2$ holds, there also holds

$$\|r(z_i) - g(x_i)\| \quad \lesssim \quad \|g(x_{\ell(n)})\|^2. \tag{4.64}$$

(iv) There is a constant $\bar{c} > 0$ such that the iterates of DIIS, applied to equation (4.47), are bounded by

$$\|z_i - z^*\| \ \leq \ \bar{c} \cdot \|x_{\ell(n)} - z^*\|. \tag{4.65}$$

Proof. The first claim follows directly from

$$r(x) - g(x) \quad = \quad g(x_n) - g(x) - J(x_n - x)$$

and the estimate (4.54). For the second inequality (4.63), note that z^* is defined as "perfect Newton update", solving (4.47); thus, there follows from (4.55) that

$$\|z^* - x^*\| \ = \ \|x_n - J^{-1}(g(x_n) - g(x^*)) - x^*\| \ \leq \ \|J^{-1}\|\frac{\gamma}{2}\|x_n - x^*\|^2 \ = \ \frac{\gamma^2}{2}\|x_n - x^*\|^2.$$

The third estimate (4.64) is concluded from (4.62), linear convergence of the algorithm and (4.56), which gives

$$\begin{aligned}
\|r(z_i) - g(x_i)\| \ &= \ \|r(z_i - x_i) + r(x_i) - g(x_i)\| \\
&\leq \ \|J(z_i - x_i)\| \ + \ \|r(x_i) - g(x_i)\| \\
&\leq \ \|J\| \ \|z_i - x_i\| \ + \ 2K\|x_i - x^*\|^2 \\
&\leq \ \|J\| \ c \cdot \|g(x_{\ell(n)})\|^2 \ + \ 2K\|x_i - x^*\|^2 \\
&\leq \ \|J\| \ c \cdot \|g(x_{\ell(n)})\|^2 \ + \ 2\rho K q^{2(i-\ell(n))}\|g(x_{\ell(n)})\|^2 \quad \lesssim \quad \|g(x_{\ell(n)})\|^2.
\end{aligned}$$

Finally, for assertion (iv) we use the relation between DIIS and GMRES iterates (4.37) to obtain

$$\begin{aligned} \|z_i - z^*\| &\leq \|v_i - z^* + r(v_i) - r(z^*)\| \\ &\leq \|I - J^{-1}\|\|Jv_i - b_n\| \leq \delta\ c_{i-\ell(n)}\ \|J\|\ \|z_{\ell(n)} - z^*\|, \end{aligned}$$

where $c_{i-\ell(n)} = \|v_i\|/\|v_{\ell(n)}\|$ is the residual reduction factor of the GMRES method for (4.47). It is known that those factors $c_{i-\ell(n)}$ form a nonincreasing sequence for any linear mapping A, see [67], and using this fact completes the proof of (iv).

□

To prove Theorem 4.12, we will have to bound the difference between the DIIS iterate x_n and iterates z_n belonging to the linear equation. The main tool used below is given in the next lemma.

Lemma 4.19. *(Difference of the Jacobian approximations)*

Let the conditions of Theorem 4.10 hold, so that the DIIS algorithm is linearly convergent, and let $\ell(n) \leq j \leq n$. Let $z_{\ell(n)}, \ldots, z_j \in E$ and let the estimate $\|x_i - z_i\| \lesssim \|g(x_{\ell(n)})\|^2$ be fulfilled for all $\ell(n) \leq i \leq j$. Moreover, suppose $\ell(i) = \ell(n)$ for all $\ell(n) \leq i \leq j$. Then, the difference between the Jacobian approximation H_j produced by the DIIS procedure applied to the equation $g(x) = 0$, and the one produced by the DIIS solver for (4.47) can be bounded by

$$\|(H_j - G_j)g(x_j)\| \leq const \cdot \|g(x_{\ell(n)})\|^2. \tag{4.66}$$

Proof. We estimate $\|H_j - G_j\|$. To this end, we use the representation (4.58) proven in Lemma 4.17; from there, because $\ell(j) = \ell(n)$, we have for H_j that

$$H_j - J^{-1} = (I - J^{-1})(I - Q_j) + \sum_{i=\ell(n)}^{j-1} \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T \bar{y}_i}$$

on the one hand; on the other hand, because $r'(x) = J$, the approximate Jacobians G_j produced by the DIIS procedure applied to the linear problem (4.47) fulfils by Lemma 4.17

$$G_j - J^{-1} = (I - J^{-1})(I - R_j), \tag{4.67}$$

with R_j denoting the projector onto

$$\text{span } \{d_i := r(z_{i+1}) - r(z_i) \mid i = \ell(n), \ldots, j-1\}.$$

Note that in the case of G_j, the latter "difference approximation" error term in (4.58) vanishes because $J^{-1}(r(z_{i+1} - r(z_i)) = z_{i+1} - z_i$ is fulfilled exactly by linear problems.

Therefore, using (4.59),

$$\begin{aligned}\|H_j - G_j\| &= \|H_j - J^{-1} - (G_j - J^{-1})\| \\ &\le \|(I - J^{-1})(R_j - Q_j)\| + \|\sum_{i=\ell(n)}^{j-1} \frac{(\bar{s}_i - J^{-1}\bar{y}_i)\bar{y}_i^T}{\bar{y}_i^T \bar{y}_i}\| \\ &\le \|R_j - Q_j\| + \alpha\|x_{\ell(n)} - x^*\|.\end{aligned}$$

We thus obtain from $\|x_{\ell(n)} - x^*\| \le \rho\|g(x_{\ell(n)})\|$ and $\|g(x_j)\| \le q^{j-\ell(n)}\rho^2\|g(x_{\ell(n)})\|$ that

$$\|(H_j - G_j)g(x_j)\| \le \|R_j - Q_j\|\,\|g(x_j)\| + \alpha\rho^3 q^{j-\ell(n)}\|g(x_{\ell(n)})\|^2,$$

so it remains to show that $\|R_j - Q_j\| \lesssim \|g(x_{\ell(n)})\|^2/\|g(x_j)\|$ to complete the proof. We prove this assertion by induction over $i = \ell(n), \ldots, j$. For $i = \ell(n)$, $R_{\ell(n)} = Q_{\ell(n)} = I$. Now, let $\|R_{i-1} - Q_{i-1}\| < c_{i-1}\|g(x_{\ell(n)})\|^2/\|g(x_{i-1})\|$ hold for some $\ell(n) < i \le j$; we then denote

$$\hat{d} := \hat{d}_{i-1} := (I - R_{i-1})(r(z_i) - r(z_{i-1})), \qquad \hat{y} := \hat{y}_{i-1} := (I - Q_{n-1})y_{i-1}$$

and use the decomposition

$$\begin{aligned}\|R_i - Q_i\| &\le \|R_{i-1} - Q_{i-1}\| + \|\frac{\hat{d}\,\hat{d}^T}{\hat{d}^T\hat{d}} - \frac{\hat{y}\,\hat{y}^T}{\hat{y}^T\hat{y}}\| \\ &\le c_{i-1}\frac{\|g(x_{\ell(n)})\|^2}{\|g(x_{i-1})\|} + \|\frac{\hat{d}\,\hat{d}^T}{\hat{d}^T\hat{d}} - \frac{\hat{y}\,\hat{y}^T}{\hat{y}^T\hat{y}}\| \\ &\le c_{i-1}\rho^2 q\frac{\|g(x_{\ell(n)})\|^2}{\|g(x_i)\|} + \|\frac{\hat{d}\,\hat{d}^T}{\hat{d}^T\hat{d}} - \frac{\hat{y}\,\hat{y}^T}{\hat{y}^T\hat{y}}\|.\end{aligned}$$

In this estimate, $\|g(x_i)\| \le q\rho^2\|g(x_{i-1})\|$, which is a conclusion from (4.56) and linear convergence, was used to obtain the last inequality. By inserting a useful zero, one sees that

$$\|\frac{\hat{d}\,\hat{d}^T}{\hat{d}^T\hat{d}} - \frac{\hat{y}\,\hat{y}^T}{\hat{y}^T\hat{y}}\| \le 2\,\|\frac{\hat{d}}{\|\hat{d}\|} - \frac{\hat{y}}{\|\hat{y}\|}\|$$

holds for the difference of the projectors. Thus, we can complete the proof by showing

$$\|\frac{\hat{d}}{\|\hat{d}\|} - \frac{\hat{y}}{\|\hat{y}\|}\| \lesssim \frac{\|g(x_{\ell(n)})\|^2}{\|g(x_j)\|}. \tag{4.68}$$

We begin by estimating with (4.64) and (4.56),

$$\begin{aligned}\|\hat{d} - \hat{y}\| &\le \|(I - R_{i-1})d - (I - R_{i-1})y\| + \|(R_{i-1} - Q_{i-1})y\| \\ &\le \|I - R_{i-1}\|\,\big(\|r(z_i) - f(x_i)\| + \|r(z_{i-1}) - f(x_{i-1})\|\big) + \|R_{i-1} - Q_{i-1}\|\,\|y_i\| \\ &\le 2C\|g(x_{\ell(n)})\|^2 + c_{i-1}\frac{\|g(x_{\ell(n)})\|^2}{\|g(x_{i-1})\|}\|g(x_i) - g(x_{i-1})\| \\ &\le \big(2C + c_{i-1}(1 + \rho^2 q)\big)\,\|g(x_{\ell(n)})\|^2 =: c_{d,y}\,\|g(x_{\ell(n)})\|^2,\end{aligned} \tag{4.69}$$

where in the last step, we have used that (4.56) together with linear convergence implies

$$\|g(x_i) - g(x_{i-1})\| \;\leq\; \|g(x_i)\| + \|g(x_{i-1})\| \;\leq\; (1+\rho^2 q)\,\|g(x_{i-1})\|.$$

We now bound the left-hand side of (4.68) by

$$\begin{aligned}
\|\frac{\hat{d}}{\|\hat{d}\|} - \frac{\hat{y}}{\|\hat{y}\|}\| &\leq \frac{\|\hat{y}-\hat{d}\|}{\|\hat{d}\|} + \|\hat{d}\|\big(\frac{1}{\|\hat{y}\|} - \frac{1}{\|\hat{y}\|}\big) \\
&= \frac{\|\hat{y}-\hat{d}\|}{\|\hat{y}\|} + \frac{\|\hat{d}\| - \|\hat{y}\|}{\|\hat{y}\|} \\
&\leq 2c_{d,y}\cdot\|g(x_{\ell(n)})\|^2\|\hat{y}\|^{-1} \;\leq\; 2\tau c_{d,y}\|g(x_{\ell(n)})\|^2\|y_i\|^{-1} \;=:\; (*),
\end{aligned}$$

where we used (4.69) to get to the last line. Finally, to estimate $\|y_i\|^{-1}$, we use that from the linear convergence of the algorithm and again, (4.56), we obtain

$$\begin{aligned}
\|y_i\| &= \|g(x_i) - g(x_{i-1})\| \;\geq\; \frac{1}{\rho}\|x_i - x_{i-1}\| \\
&\geq \frac{1-q}{\rho}\|x_{i-1} - x^*\| \;\geq\; \frac{(1-q)q^{-(j-(i-1))}}{\rho^2}\|g(x_j)\|;
\end{aligned}$$

thus, using linear convergence and (4.56) once more, we get

$$(*) \;\leq\; \frac{2\tau c_{d,y}\rho^2 q^{j-(i-1)}}{(1-q)}\,\frac{\|g(x_{\ell(n)})\|^2}{\|g(x_j)\|} \;=:\; c\cdot\frac{\|g(x_{\ell(n)})\|^2}{\|g(x_j)\|}.$$

This proves (4.68) and thus the assertion.

□

We are now prepared to prove Theorem 4.12.

Proof of Theorem 4.12. We start by choosing the constants $\epsilon, \delta > 0$ as in Theorem 4.10, so that we can assume linear convergence of the algorithm. If necessary, we decrease ϵ such that for the constants γ from (4.63) and $\bar{c}$ from (4.65), the ball with radius $r := \bar{c}\epsilon + \gamma^2\epsilon^2/2$ lies in E. We now fix $n \in \mathbb{N}$ and set $\ell := \ell(n)$ for brevity. We decompose the error into the three terms

$$\|g(x_{n+1})\| \;\leq\; \underbrace{\|g(z^*)\|}_{\text{(I)}} + \underbrace{\|g(z_{n+1}) - g(z^*)\|}_{\text{(II)}} + \underbrace{\|g(x_{n+1}) - g(z_{n+1})\|}_{\text{(III)}} \tag{4.70}$$

with the quantities z_{n+1}, z^* from Definition 4.11. We will see that estimation of the single terms will then give the three error components of the estimate (4.48).
For the first term, we obtain from Lemma 4.18(ii) that

$$\text{(I)} \;=\; \|g(z^*) - g(x^*)\| \;\leq\; \rho\|z^* - x^*\| \;\leq\; \frac{\rho\gamma^2}{2}\|x_n - x^*\|^2 \;\leq\; \frac{\rho^2\gamma^2}{2}\|g(x_n)\|^2,$$

and thus the first part of the estimate (4.48).
We continue with estimation of (II). By the choice of ϵ, and with Lemma 4.18(ii),(iv) we obtain for all $i \in \ell, \dots, n+1$ that

$$\|z_i - x^*\| \;\le\; \|z_i - z^*\| \;+\; \|z^* - x^*\| \;\le\; \bar{c} \cdot \|x_\ell - x^*\| + \frac{\gamma^2}{2}\|x_n - x^*\|^2 \;<\; r,$$

so that $z_i \in E$. From (4.53) and $\|z_{n+1} - z^*\| \le \gamma \|J(z_{n+1} - z^*)\|$, we thus get

$$\begin{aligned}\|g(z_{n+1}) - g(z^*)\| \;\le\; & \|J(z_{n+1} - z^*)\| \;+\; 2K\, \|z_{n+1} - z^*\|\, \max\{\|z_{n+1} - x^*\|, \|z^* - x^*\|\} \\ & \|J(z_{n+1} - z^*)\| \cdot (1 + 2K\gamma\, \max\{\|z_{n+1} - x^*\|, \|z^* - x^*\|\}),\end{aligned}$$

and now estimate both factors of the last line separately. The main point for estimation of $\|J(z_{n+1} - z^*)\|$ is that $J(z_{n+1} - z^*) = r(z_{n+1})$, the residual of the "virtual" DIIS/GMRES procedure as defined in 4.11. Therefore, we can estimate this term with the help of Theorem 4.7, with (4.62) and with (4.56) by

$$\begin{aligned}\|J(z_{n+1} - z^*)\| \;=\; \|r(z_{n+1})\| \;&\le\; d_{n-\ell}\;\; \|r(x_\ell)\| \\ &\le\; d_{n-\ell}\;\; (\|g(x_\ell)\| \;+\; 2K\|x_\ell - x^*\|^2) \\ &\le\; d_{n-\ell}\;\; (1 + 2\rho K\epsilon)\;\; \|g(x_\ell)\|. \qquad (4.71)\end{aligned}$$

For the second factor, we obtain by using $\|z_{n+1} - x^*\| \le \|z_{n+1} - z^*\| + \|z^* - x^*\|$, (4.47) and (4.63) that

$$\max\{\|z_{n+1} - x^*\|, \|z^* - x^*\|\} \;\le\; \|z_{n+1} - z^*\| + \|z^* - x^*\| \;\le\; \gamma\|r(z_{n+1})\| + \frac{\gamma^2}{2}\|x_n - x^*\|^2.$$

We let $\epsilon := \|x_\ell - x^*\|$ as before and use (4.71), (4.56) and linear convergence to get

$$\begin{aligned}\gamma\|r(z_{n+1})\| + \frac{\gamma^2}{2}\|x_n - x^*\| \;&\le\; \gamma\; d_{n-\ell}\;\; (1 + 2\rho K\epsilon)\;\; \|g(x_\ell)\| + \frac{\gamma^2}{2}\|x_n - x^*\|^2 \\ &\le\; \big(\gamma\; \rho\; (1 + 2\rho K\epsilon) + \frac{\gamma^2}{2} q^{2(n-\ell)}\epsilon\big)\epsilon \;\;=:\;\; \omega\epsilon;\end{aligned}$$

where we have used that $d_{n-\ell} \le 1$, see the proof of (4.65). Altogether, (II) can now be bounded by

$$\begin{aligned}\text{(II)} \;=\; \|g(z_{n+1}) - g(z^*)\| \;&\le\; d_{n-\ell}\; (1 + 2\rho K\epsilon)(1 + 1\kappa\gamma\omega\epsilon)\; \|g(x_\ell)\| \\ &\le\; d_{n-\ell}\; (1 + c_2 \cdot \epsilon)\; \|g(x_\ell)\|\end{aligned}$$

with c_2 suitably chosen; this gives the second term of the estimate (4.48). The third part of (4.70) can be estimated by

$$\text{(III)} \;=\; \|g(x_{n+1}) - g(z_{n+1})\| \;\le\; \rho\; \|x_{n+1} - z_{n+1}\|;$$

to complete the proof of (4.48), we now show by induction that for all $i = \ell, \dots, n+1$, $\|x_i - z_i\| \lesssim \|g(x_\ell)\|^2$. For $x_\ell = z_\ell$, there is nothing to show. For the induction step, we fix $i \in \{\ell, \dots, n\}$ and note that by recursively using the definition of the iterates,

$$x_{i+1} - z_{i+1} \;=\; x_\ell - z_\ell \;+\; \sum_{j=\ell}^{i} G_j r(z_j) - H_j g(x_j).$$

Therefore, because $x_\ell = z_\ell$,

$$\|x_{i+1} - z_{i+1}\| \quad \leq \quad \sum_{j=\ell}^{i} \|H_j g(x_j) - G_j r(z_j)\|. \tag{4.72}$$

To estimate the terms to the right, we note at first that for all $j = \ell, \ldots, i$, using (4.56), the induction hypothesis and Lemma 4.18(iii), there holds

$$\begin{aligned} \|g(x_j) - r(z_j)\| &\leq \|g(x_j) - g(z_j)\| + \|g(x_j) - r(z_j)\| \\ &\leq \rho\|x_j - z_j\| + \text{const}\|g(x_j)\|^2 \quad \leq \quad \kappa\|g(x_\ell)\|^2 \end{aligned}$$

for a suitable constant $\kappa > 0$. Additionally, we observe that from (4.67), there follows $\|G_i\| \leq c$ for some $c > 0$ and all $i = \ell, \ldots, n$. We now estimate the terms in (4.72) separately: Using the previous remarks and Lemma 4.19, we get for each $i = \ell, \ldots, n$ that

$$\|H_i g(x_i) - G_i r(z_i)\| \quad \leq \quad \|(H_i - G_i)g(x_i)\| + \|G_i\| \, \|g(z_i) - r(z_i)\| \quad \lesssim \quad \|g(x_\ell)\|^2.$$

This completes the proof of Theorem 4.12.

□

4.5 Concluding remarks on Section 4

We have identified DIIS with the reverse, projected Broyden's method given by formula (4.9). By this connection between DIIS and the family of Broyden-like methods and Krylov space methods, the development of new problem-adapted variants of DIIS and related convergence accelerators may well profit from the theoretical as well as from the practical experience made with Broyden-like methods and Newton-Krylov type methods. Results for the reverse, projected Broyden's method corresponding to DIIS [82, 102] show that although DIIS is applied to quantum chemical problems with great success, it seems to be inferior to the according "forward" method.
On the other hand, DIIS provides a great amount of examples where the projected backward method works fairly well in practice (and where it may be preferred due to its simple implementation in the form of DIIS). This is in particular the case when DIIS is applied to linear equations, for which we showed that the convergence of DIIS is fixed by the mostly favourable convergence behaviour of the according GMRES procedure. DIIS, applied to nonlinear equations, can therefore be viewed as a globalization of GMRES to nonlinear equations, and we have shown that the convergence behaviour is still related to the properties of a linear equation for the Jacobian J if nonlinear effects are small. This is in agreement with similar results for Broyden's method [93], and we conjecture that if nonlinearities are mild, DIIS still shares the favourable properties of the according GMRES procedure. It would be interesting to theoretically and practically investigate the influence of increasing nonlinearities on the performance of the DIIS solver further in the future.

□

Conclusion and outlook

In this work, we have analysed aspects of some of the most widely applied methods of quantum chemistry. To apply mathematical concepts that are common in the context of partial differential equations and operator eigenvalue problems, we have, in contrast to the normal proceeding in the literature concerned with quantum chemistry, insisted that the electronic eigenvalue problem and also the approximation methods analysed, i.e. Hartree-Fock, DFT and CI (Section 2) as well as the CC method (Section 3) be formulated in the suitable original, infinite-dimensional spaces dictated by the axioms and framework of mathematical physics. In particular in the context of the Coupled Cluster method, some technical difficulties had to be overcome to obtain the according infinite dimensional formulation, the continuous Coupled Cluster method.

Nevertheless, this proceeding has put us in a position from which we could show that also in the respective suitable infinite-dimensional (mostly Sobolev) spaces, the operators under consideration fulfil the assumptions necessary for the then more or less straightforward application of functional analyical concepts. Our approach has thus rewarded us with the ability to derive concepts otherwise unattainable: The results of our convergence analysis for some of the main algorithms used of quantum chemistry hold for the methods and algorithms formulated in the continuous space on the one hand, thus providing a solid basis for an adaptive treatment, cf. the remarks in Section 1.7, on the other, they hold as well for the according discretized methods, where the estimates are uniform with respect to the discretization parameters. Additional results obtained were the goal-oriented error estimators and quasi-optimality results for the Coupled Cluster method (Theorems 3.21, 3.24). For the DIIS method analysed in Section 4, the establishment of connections to methods well-known in numerical mathematics has enabled us to obtain some convergence results from those for GMRES and quasi-Newton methods.
The results proven should now be used as a theoretical basis to implement goal-oriented error estimators and to analyse adaptive algorithms set on top of the successfully applied existing practical methods of quantum chemistry. Also, our convergence results should be extended to further methods of quantum chemistry, as for instance the multi-reference version of the Coupled Cluster method; the analysis given here may also serve as a basis to obtain theoretical results for linear-scaling methods, both in the context of density functional theory and of Coupled Cluster theory.

To conclude, allow the author to express the hope that this work has made a useful contribution to the ambition of numerical analysis to bridge the gap between the theoretical investigation of the properties of quantum mechanical Hamiltonians and the electronic Schrödinger equation on the one side, and the methods and algorithms used in practical applications in the fields of quantum chemistry and electronic structure calculation on the other. Hopefully, the results provided in this work can serve to stimulate the further intertwinement of the scientific communities involved in the theoretical investigation and in the practical treatment of the electronic Schrödinger equation.

Notation

This list of symbols features an overwiew of the most important re-occurring notations of this work. Within the respective assortings, the overview features an alphabetical order; Greece letters are sorted in by the first letter of their English transcription.

Spaces & manifolds:

$\mathbb{F}$	Fock space, p. 22
$\mathcal{G}$	Grassmann manifold over V^N, p. 42
$L^2(\Omega)$	Space of complex-valued, measurable, square-integrable functions defined on a measure space Ω, p. 3
$\mathcal{L}^2 = \mathcal{L}^2_N$	N-fold tensor space of $L^2(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$, p.3
$\widehat{\mathcal{L}}^2$	Space of antisymmetric functions from $\mathcal{L}^2$, p.10
$\mathcal{L}^2_k$	Eigenspace of the z-spin operator corresponding to eigenvalue $s_k = -\frac{N}{2} + k$, p.11
$\mathcal{L}^2_{\mathbb{R}}, \mathcal{L}^2_{\mathbb{C}}$	Space of purely real-valued/purely imaginary-valued wave functions from $\mathcal{L}^2$, p.12
$\mathbb{L}^2 = \mathbb{L}^2_N$	Space of real-valued, antisymmetric functions from $\mathcal{L}^2$, p.12
$\mathbb{L}^2_k$	Space of real valued, antisymmetric functions from $\mathcal{L}^2_k$, p.12
H^1	Space of real valued, antisymmetric functions from $\mathcal{H}^t_k$ ($= \mathbb{H}^t_k$, with k fixed later)
$H^1(\Omega)$	Sobolev subspace of $L^2(\Omega)$ of one time weakly differentiable functions, defined on a measure space Ω, p. 6
$\mathcal{H}^t$	Abbreviation for the Sobolev space $H^t(\mathbb{R}^{3N} \times \Sigma^N) \subseteq \mathcal{L}$, p.6
$\mathcal{H}^t_{\otimes}$	N-fold tensor product space of $H^1(R^3 \times \Sigma)$, p.6
$\widehat{\mathcal{H}}^t$	Space of antisymmetric $\mathcal{H}^t$-functions, p.10
$\mathcal{H}^t_k$	Space of functions from $\mathcal{L}^2_k$ with Sobolev regularity t, p.11
$\mathbb{H}^t_k$	Space of real valued, antisymmetric functions from $\mathcal{H}^t_k$, p.12
$\widehat{\mathcal{H}}^{-t}, \mathbb{H}^{-t}, \ldots$	Dual spaces of $\widehat{\mathcal{H}}^t, \mathbb{H}^t, \ldots$
$\mathcal{T}(\mathcal{G})$	Tangent space of the Grassmann manifold $\mathcal{G}$ at $[\Phi] \in \mathcal{V}$, p. 43
V	General Hilbert space, in Sec. 2 belonging to a Gelfand triple (see p. 41).
$\mathbb{V}$	Coefficient space used in CC calculations, p. 72
$\mathcal{V}$	Stiefel manifold over V^N, p. 42
X	In Sec. 2: Shorthand notation for some $L^2(\Omega, \mathbb{R})$ or $L^2(\Omega, \mathbb{C})$, p. 41.

ll

Operators:

$\mathcal{A}, \mathcal{B}, \ldots$	Expansion of $A, B, \ldots : V \to V'$ to an operator $V^N \to (V')^N$, p. 41
$A_N, F_N, \ldots$	N-fold canonical Kronecker product of an operator $A, F, \ldots$, p. 5
a_f	Annihilation operator for f, p. 23
a_P	Annihilation operator for χ_P from one particle basis, p. 23
$a_f^\dagger$	Creation operator for f, p. 23
$a_P^\dagger$	Creation operator for χ_P from one particle basis, p. 23
D^*	In Sec. 2: L^2-projector on span$[\Phi_0]$.
F^{HF}	Fock operator, p. 45
F^{KS}	Kohn-Sham operator, p. 45
f	Coupled Cluster function, p. 75
g	Function subject to the root problem (4.1) in Section 4 lll
H	Electronic Hamiltonian, p. 8
$\widehat{H}$	Second quantization (weak) Hamiltonian, p. 26
h	Bilinear form induced by H; $h : \mathbb{H}^1 \times \mathbb{H}^1 \to \mathbb{R}$, p. 14
P	In Sec. 4: Preconditioning mapping, p. 95
$\mathcal{P}^a$	Antisymmetrization projector on $\mathbb{L}$, p. 9
$\mathcal{Q}$	Isomorphism from $\mathbb{L}'_k$ onto $\mathbb{L}_k$, p. 19
$S, T, \ldots$	In Sec. 3: Cluster operators, p. 64
S_N^z	z-spin operator for N-electron systems, p. 11
$\mathbf{U}$	Unitary $\mathbb{R}^{N\times N}$-matrix
$X_\mu = X^{A_1,\ldots,A_r}_{I_1,\ldots,I_r}$	Excitation operator, mapping Ψ_0 to $\Psi_\mu = \Psi^{A_1,\ldots,A_r}_{I_1,\ldots,I_r}$, p. 62/p. 63

Numbers & indices:

$A,\ B,\ C$	In Section 3: Indices corresponding to occupied orbitals
$\alpha \oplus \beta, \alpha \ominus \beta$	Index operations, p. 62
D	Section 3: Dimension of truncated one particle basis set Section 4: Dimension of discretized space
E_0	Electronic ground state energy of the molecule, p. 14
E^*	Ground state energy of electronic configuration of spin $k \in \{0, \ldots, N\}$, p. 15
$I,\ J,\ K$	In Section 3: Indices corresponding to virtual orbitals
λ_i	Eigenvalue of the Fock/Kohn-Sham operator
Λ_0	Sum of k eigenvalues of the Fock/Kohn-Sham operator, p. 30
N	Number of electrons, equals number of occupied orbitals
$\overline{p}, \underline{q}, \ldots$	Indices labelling "spin up" resp. "spin down" orbitals, p. 18
$P,\ Q,\ R,\ S$	Indices labelling spin orbitals, containing number and spin, p. 18
$r(\mu)$	Rank of a Slater determinant, p. 62
V	Number of virtual orbitals in truncated basis set, $V = D - N$.

Functions & vectors:

$\|\cdot\|$	Canonical norm on $L^2(\Omega)$
$\|\cdot\|_1$	Canonical norm on $H^1(\Omega)$
χ_P	Spin orbitals/spin basis functions for $H^1(\mathbb{R}^3 \times \{\pm\frac{1}{2}\})$, p. 18
f	Coupled Cluster function, p. 75
g	Function subject to the root problem (4.1) in Section 4
$\mathcal{J}$	Functional subject to constraint minimization in Sec. 2
$\varphi, \psi, \ldots$	Spin free/spatial one particle functions, i.e. functions from $H^1(\mathbb{R}^3, \mathbb{R})$
φ_p	Functions from spin free one particle basis, p. 18
Φ	Sec. 2: Vector from V^N, $\Phi = (\varphi_1, \ldots, \varphi_N)$
$\Psi, \Psi', \ldots$	Functions from the tensor product space $\mathcal{L}$
Ψ_μ	Basis function from the tensor basis $\mathbb{B}$ resp. $\mathbb{B}_k$, p. 19
Ψ_0	Sec. 2: Hartree-Fock solution, p. 36
	Sec. 3: Reference Slater determinant, p. 60
$\underline{\Psi}$	Solution of the electronic Schrödinger equation, p. 15
Ψ^*	Sec. 3: Correlation correction, $\underline{\Psi} - \Psi_0$, p. 64
t, s	Coefficient vectors from the Coupled Cluster coefficient space $\mathbb{V}$
t^*	Solution of the Coupled Cluster equations, p. 73

Sets:

B	Basis of spatial orbitals, p. 18
B^Σ	Basis of spin orbitals, p. 18
$\mathcal{B}$	Tensor basis of $\mathcal{H}^1$, p. 18
$\mathbb{B}$	Tensor basis of $\mathbb{H}^1$, p. 19
$\mathbb{B}_k$	Tensor basis of $\mathbb{H}^1_k$, p. 19
$\mathcal{I}$	Index set containing indices for spin orbitals, p. 18
$N]$	The set $\{i \in \mathbb{N} \mid 1 \leq i \leq N\}$
$\mathcal{M}$	Index set containing indices for tensor basis functions $\Phi_\mu \in \mathbb{B}$, p. 21
$\mathcal{M}_k$	Index set containing indices for tensor basis functions $\Phi_\mu \in \mathbb{B}_k$, p. 21
$\mathcal{M}^*, \mathcal{M}^*_k$	$\mathcal{M}^* := \mathcal{M} \backslash \{\mu_0\}$, $\mathcal{M}^*_k := \mathcal{M}_k \backslash \{\mu_0\}$, p. 62
$\mathrm{spin}(N)$	Set of possible z-spins for an N-electron system, p. 11
Σ^N	Set of possible spin vectors σ, p. 3
$S(N)$	Group of permutations on N elements
$\mathrm{spec}(A)$	Spectrum of an operator A, p. 13
occ	Sets of indices belonging to occupied orbitals, p. 61
virt	Sets of indices belonging to occupied orbitals, p. 61

References

[1] ABINIT is a common project of the Université Catholique de Louvain, Corning Incorporated, and other contributors, for further details see http://www.abinit.org

[2] P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, 2007.

[3] P.-A. Absil, R. Mahony, R. Sepulchre, *Riemannian geometry of Grassmann manifolds with a view on algorithmic computation*

[4] S. Agmon, *Lectures on exponential decay of solutions of second-order elliptic equations*, Princeton University press, Princeton, 1982.

[5] D. C. Allen, T. A. Arias, J. D. Joannopoulos, M. C. Payne, M. P. Teter, *Iterative minimization techniques for ab initio total energy calculation: molecular dynamics and conjugate gradients*, Rev. Modern Phys. 64, 4, p. 1045, 1992.

[6] H. W. Alt, *Lineare Funktionalanalysis,* 5. Auflage, Springer, Berlin, 2006.

[7] A. Anantharaman, E. Cancès, *Existence of minimizers for Kohn-Sham models in quantum chemistry*, Annales de l'Institut Poincare, Non Linear Analysis, 26, 6, p. 2425, 2009.

[8] H. Araki, *Einführung in die Axiomatische Quantentheorie, Vorlesung an der ETH Zürich, Wintersemester 1961/62*, available at TU Berlin, 1962.

[9] T. A. Arias, *Multiresolution analysis of electronic structure: semicardinal and wavelet bases*, Rev. Mod. Phys. 71, p. 267, 1999.

[10] T. A. Arias, A. Edelman, S. T. Smith, *The Geometry of Algorithms with Orthogonality Constraints*, SIAM J. Matrix Anal. and Appl. 20, 2, p. 303, 1999.

[11] M. Arioli, Vlastimil Pták, Zdenek Strakoš, *Krylov sequences of maximal length and convergence of GMRES*, BIT Numerical Mathematics 38, 4, p. 636, 1998.

[12] A. A. Auer, G. Baumgärtner et al., *Automatic Code Generation for Many-Body Electronic Structure Methods: The tensor contraction engine*, Molecular Physics 104, 2, p. 211, 2006.

[13] A. A. Auer, M. Nooijen, *Dynamically screened local correlation method using enveloping localized orbitals*, J. Chem. Phys. 125, 24104, 2006.

[14] P. Y. Ayala, G. E. Scuseria, *Electronic correlation in large molecular systems using the atomic orbital formalism. The case of intermolecular interactions in chrystalline urea as an example*, J. Comput. Chem. 21, p. 1524, 2000.

[15] V. Bach, E. H. Lieb, M. Loss, J. P Solovej, *There are no unfilled shells in unrestricted Hartree-Fock theory.* Phys. Rev. Letters, 72, 19, p. 2981, 1994.

[16] W. Bangerth, R. Rannacher, *Adaptive finite element methods for differential equations*, Birkhäuser, 2003.

[17] R. J. Bartlett, M. Musial, *Coupled-cluster theory in quantum chemistry.* Rev. Mod. Phys., 79, p. 291, 2007.

[18] R. J. Bartlett, G. D. Purvis, *Many-body perturbation theory, coupled-pair many-electron theory, and the importance of quadruple excitations for the correlation problem*, Int. J. Quantum Chem. 14, p. 561, 1978.

[19] R. J. Bartlett, *Many-body perturbation theory and coupled cluster theory for electronic correlation in molecules*, Ann. Rev. of Phys. Chem. 32, p. 359, 1981.

[20] H. Bauer, *Maß - und Integrationstheorie*, de Gruyter Lehrbücher, 2. Aufl., 1992.

[21] T. L. Beck, *Real-space mesh techniques in density-functional theory*, Rev. Mod. Phys. 72, p. 1041, 2000.

[22] R. Becker, R. Rannacher, *An optimal control approach to error estimation and mesh adaptation in finite element methods*, Acta Numerica 2000 (A. Iserles, ed.), p. 1, Cambridge University Press, 2001.

[23] B. Beckermann, A. B. J. Kuijlaars, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal. 39, p. 300, 2001.

[24] B. Beckermann, A. B. J. Kuijlaars, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal. 14, p. 1, 2002.

[25] R. E. Bellman, *Adaptive Control Processes.* Princeton University Press, Princeton, 1961.

[26] U. Benedikt, M. Espig, W. Hackbusch, A. A. Auer, *A new Approach for Tensor Decomposition in Electronic Structure Theory*, to be submitted.

[27] F. A. Berezin, *The Method of Second Quantization*, Academic Press. 1966.

[28] F. A. Berezin, M. A. Shubin, *The Schrödinger equation*, Kluwer Academic Publishing, 1991.

[29] D. E. Bernholdt, R. J. Bartlett, *A Critical Assessment of Multireference-Fock Space CCSD and Perturbative Third-Order Triples Approximations for Photoelectron Spectra and Quasidegenerate Potential Energy Surfaces*, Advances in Quantum Chemistry 34, p. 261, 1999.

[30] bigDFT, http://www-drfmc.cea.fr/sp2m/L_Sim/BigDFT/index.en.html

[31] R. F. Bishop, *An overview of coupled cluster theory and its applications in physics*, Theor Chim Acta 80, p. 95, 1991.

[32] M. Born, R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Ann. Phys. 389, 20, p. 457, 1927.

[33] S. F. Boys, *Construction of some molecular orbitals to be approximately invariant for changes from one molecule to another*, Rev. Mod. Phys. 32, p. 296, 1960.

[34] J. H. Bramble, J. E. Pasciak, and A. V. Knyazev, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math. 6, p. 159, 1996.

[35] C.G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Mathematics of Computation 19, p. 577, 1965.

[36] C. G. Broyden, J. E. Dennis, J. J. Moré, *On the local and superlinear convergence of Quasi-Newton methods*, J. Inst. Math. Appl. 12, p. 223, 1973.

[37] R. J. Buenker, S. D. Peyerimhoff, *Individualized configuration selection in CI calculations with subsequent energy extrapolation*, Theoret. Chim. Acta 35, p. 33, 1974.

[38] H. J. Bungartz, M. Griebel, *Sparse Grids*, Acta Numerica, p. 1, Cambridge Univ. Press, 2004.

[39] C. F. Bunge, *Scalable implementation of analytic gradients for second-order Z-averaged perturbation theory using the distributed data interface*, J. Chem. Phys. 124, 14107, 2006.

[40] C. F. Bunge, R. Carbó-Dorca, *Selected configuration interaction with truncation energy error and application to the Ne atom*, J. Chem. Phys. 125, 14108, 2006.

[41] E. J. Bylaska, W. A. de Jong et al., *NWChem, A Computational Chemistry Package for Parallel Computers, Version 5.1.* Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA. A modified version, 2007.

[42] E. Cancès, C. Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, M2AN 34, p. 749, 2000.

[43] E. Cancès, C. Le Bris, Y. Maday, *Méthodes mathématiques en chimie quantique*, Springer, 2006.

[44] A. Chamorro, *Method for construction of operators in Fock space*, Pramana 10, p. 83, 1978.

[45] S. R. Chinnamsetty, M. Espig, B. N. Khoromskij, W. Hackbusch, H.-J. Flad, *Tensor product approximation with optimal rank in quantum chemistry*, J. Chem. Phys. 127, 084110, 2007.

[46] O. Christiansen, *Coupled cluster theory with emphasis on selected new developments*, Theor. Chem. Acc. 116, p. 106, 2006.

[47] P. G. Ciarlet (Editor), J. L. Lions, *Handbook of Numerical Analysis, Volume II: Finite Element Methods (Part I)*, Elsevier, 1991.

[48] P. G. Ciarlet (Editor), C. Lebris (Guest Editor), *Handbook of Numerical Analysis, Volume X: Special Volume. Computational Chemistry.* Elsevier, 2003.

[49] J. Čížek, *Origins of coupled cluster technique for atoms and molecules*, Theor. Chim. Acta 80, p. 91, 1991.

[50] F. Coester, *Bound states of a many-particle system*, Nucl. Phys. 7, p. 421, 1958.

[51] F. Coester, H. Kümmel, *Short range correlations in nuclear wave functions*, Nucl. Phys. 17, p. 477, 1960.

[52] A. Cohen, W. Dahmen, R. DeVore, *Adaptive wavelet methods for elliptic operator equations: Convergence rates.*, Math. Comp. 70, p. 27, 2001.

[53] A. Cohen, W. Dahmen, R. DeVore, *Adaptive wavelet methods II: Beyond the elliptic case*, Found. Comput. Math 2, p. 2002, 2000.

[54] A. Cohen, W. Dahmen, R. DeVore, *Adaptive wavelet schemes for nonlinear variational problems*, SIAM J. Numer. Anal. 41, 5, p. 1785, 2003.

[55] *Computational Chemistry Comparison and Benchmark Data Base, National Institute of Standards and Technology*, www.cccbdb.nist.org.

[56] T. D. Crawford, H. F. Schaeffer III, *An introduction to coupled cluster theory for computational chemists*, Reviews in Computational Chemistry 14, p. 33, 2000.

[57] P. Császár, P. Pulay, *Geometry optimization by direct inversion in the iterative subspace*, J. of Molecular Structure 114, p. 31, 1984.

[58] W. Dahmen, T. Rohwedder, R. Schneider, A. Zeiser, *Adaptive Eigenvalue Computation - Complexity Estimates*, Num. Math. 110, 3, p. 277, 2008.

[59] M. S. Daw, *Model for energetics of solids based on the density matrix*, Phys. Rev. B 47, p. 10895, 1993.

[60] J. E. Dennis Jr., R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[61] P. Deuflhard, R. Freund, A. Waltera, *Fast secant methods for the iterative solution of large nonsymmetric linear systems*, Impact of Computing in Science and Engineering 2, 3, p. 2446, 1990.

[62] P. A. M. Dirac, *Quantum Mechanics of Many-Electron Systems*, Proceedings of the Royal Society of London, Series A, Vol. CXXIII, pp 714, 1929.

[63] R. M. Dreizler, E. K. U. Gross, *Density functional theory*, Springer, 1990.

[64] N. Dunford, J.T. Schwartz, *Linear operators, Part II*, Interscience Publishers, John Wiley & Sons, 1961.

[65] E. G. D'yakonov, *Optimization in solving elliptic problems*, CRC Press, 1996.

[66] F. Eckert, P. Pulay, and H.-J. Werner, *Ab Initio Geometry Optimization for Large Molecules*, J. Comp. Chem. 18, p. 1473, 1997.

[67] S. C. Eisenstat, H. C. Elman, M. H. Schultz, *Variational iterative methods for non-symmetric systems of linear equations*, SIAM J. Numer. Anal. 20, p. 345, 1983.

[68] E. Emmrich, *Gewöhnliche und Operator-Differentialgleichungen*, Vieweg, 2004.

[69] B. Engels, *A detailed study of the configuration selected multireference configuration interaction method combined with perturbation theory to correct the wave function*, J. Chem. Phys. 100, p. 1380, 1994.

[70] M. Espig, W. Hackbusch, T. Rohwedder, R. Schneider, *Variational Calculus with Sums of Elementary Tensors of Fixed Rank*, submitted to Num. Math.

[71] S. Evangelisti, J. P. Daudey, J. P. Malrieu, *Convergence of an improved CIPSI algorithm*, Chem. Phys. 75, p. 91, 1983.

[72] H. Everett, *Relative State Formulation of Quantum Mechanics*, Reviews of Modern Physics 29, p. 454, 1957.

[73] R. P. Feynman, *Quantum Electrodynamics*, Basic Books, 1961.

[74] R. P. Feynman, R. Leighton, M. Sands, *Lectures in Physics*, 3 volumes, Addison-Wesley Longman, 1998.

[75] T. H. Fischer, J. Almlöf, *General methods for geometry and wave function optimization*, J. Phys. Chem. 92, p. 9768, 1992.

[76] H.-J. Flad, T. Rohwedder, R. Schneider, *Adaptive methods in quantum chemistry*, Z. Phys. Chem. 224, 3-4, p.651, 2010.

[77] V. Fock, *Konfigurationsraum und zweite Quantelung*, Z. Phys. 75, p. 622, 1932.

[78] W. C. M. Foulkes, L. Mitas, R. J. Needs, G. Rajagopal, *Quantum Monte Carlo simulations of solids,* Rev. Mod. Phys. 73, p. 33, 2001.

[79] S. Fournais, M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and T. Østergaard Sørensen, *Sharp regularity results for Coulombic many-electron wave functions*, Commun. Math. Phys. 255, p. 183, 2005.

[80] H. Gajewski, K. Gröger, K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie Verlag, 1974.

[81] D. M. Gay, *Some convergence properties of Broyden's method*, SIAM J. Numer. Anal. 16, p. 623, 1979.

[82] D. M. Gay, R. B. Schnabel, *Solving systems of nonlinear equations by Broyden's method with projected updates*, Nonlinear Programming 3, Academic Press, 1978.

[83] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, R. Schneider, *Daubechies wavelets as a basis set for density functional pseudopotential calculations*, J. Chem. Phys., 129, 014109, 2009.

[84] C. Gerthsen (author), D. Meschede (publisher), *Physik*, Springer Berlin, 2002.

[85] D. Gilbarg, N. S. Trudinger, *Elliptic partial differential equations of second order*, Grundlehren der mathematischen Wissenschaft 224, Springer, 1998.

[86] S. Goedecker, *Wavelets and their Application for the Solution of Partial Differential Equation*, Presses Polytechniques Universitaires et Romandes, Lausanne, 1998.

[87] S. Goedecker, *Linear scaling electronic structure methods*, Reviews of Modern Physics 71, 4, p. 1085, 1999.

[88] S. Goedecker and L. Colombo, *Efficient Linear Scaling Algorithm for Tight-Binding Molecular Dynamics,* Phys. Rev. Lett. 73, p. 122, 1994.

[89] X. Gonze et al., *First-principles computation of material properties: the ABINIT software project*, Comput. Mater. Sci. 25 , p. 478, 2002.

[90] X. Gonze et al., *A brief introduction to the ABINIT software package*, Zeit. Kristallogr. 220, p. 558, 2005.

[91] G.H. Golub, L.-Z. Liao, Continuous methods for extreme and interior eigenvalue problems, Linear Algebra Appl. 415, p.31, 2006.

[92] M. Griebel, J. Hamaekers, *Sparse grids for the Schrödinger equation.* ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique 41, 2, 215, 2007.

[93] A. Griewank, *The local convergence of Broyden-like Methods on Lipschitzian problems in Hilbert spaces*, SIAM Num. Anal. 24, 3, p. 684, 1987.

[94] W. H. Greub, *Multilinear Algebra*, Springer, 1967.

[95] W. Hackbusch, *Elliptic differential equations - Theory and numerical treatment*, Springer-Verlag, 2007.

[96] W. Hackbusch, *Iterative solution of large sparse systems of equations*, Springer, 1994.

[97] G. A. Hagedorn, A. Joye, *Mathematical analysis of Born-Oppenheimer approximations*, Proceedings of symposia in pure mathematics 76, 1, p. 203, 2007.

[98] E. Hairer, C. Lubich, G. Wanner, *Geometrical Numerical Integration - Structure-Preserving Algorithms for Ordinary Differential Equations*, 2nd edition, Springer, 2006.

[99] M. Hanrath and B. Engels, *New algorithms for an individually selecting MR-CI program*, Chem. Phys. 225, p. 197, 1997.

[100] C. Hampel, H.-J. Werner, *Local treatment of electron correlation in coupled cluster theory*, J. Chem. Phys. 104, p. 6286, 1996.

[101] R. J. Harrison, *Approximating full configuration interaction with selected configuration interaction and perturbation theory*, J. Chem. Phys. 94, p. 5021, 1991.

[102] R. J. Harrison, *Krylov Subspace Accelerated Inexact Newton Method for Linear and Nonlinear Equations*, J. Comp. Chem. 25, p. 328, 2003.

[103] T. Helgaker, P. Jørgensen, J. Olsen, *Molecular Electronic-Structure Theory*, John Wiley & Sons, 2000.

[104] S. Hirata, *Tensor contraction engine: Abstraction and automated parallel implementation of Configuration-Interaction, Coupled-Cluster, and Many-Body perturbation theories*, J. Phys. Chem. A, 46, p. 9887, 2003.

[105] P. D. Hislop, I. M. Sigal, *Introduction to spectral theory with application to Schrödinger operators*, Appl. math. sc. 113, Springer, 1996.

[106] M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, *Local properties of solutions of Schrödinger equations*, Commun. Partial Diff. Eq. 17, p. 491, 1992.

[107] M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, H. Stremnitzer, *Local properties of Coulombic wave functions*, Commun. Math. Phys. 163, p. 185,1994.

[108] M. Hoffmann-Ostenhof, R. Seiler, *Cusp conditions for eigenfunctions of n-electron systems*, Phys. Rev. A 23 p. 21, 1981.

[109] P. Hohenberg, W. Kohn, *Inhomogeneous Electron Gas*, Phys. Rev. 136, p. 864, 1964.

[110] Y. Huang, H. van der Vorst, *Some Observations on the Convergence Behavior of GMRES*, Tech. Rep. 89-09, Faculty of Technical Mathematics and Informatics, Delft University of Technology, The Netherlands, 1989.

[111] W. Hunziker, I. M. Sigal, *The quantum N-body problem*, J. Math. Phys., Vol. 41, 6, 2000.

[112] B. Huron, J. P. Malrieu, P. Rancurel, *Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions*, J. Chem. Phys. 58, p. 5745, 1973.

[113] T. Kato, *Fundamental properties of Hamiltonian operators of Schrödinger type*, Trans. Am. Math. Soc. 70, p. 195, 1951.

[114] T. Kato, *On the eigenfunctions of many-particle systems in quantum mechanics*, Comm. on pure and applied mathematics X, p. 151, 1957.

[115] M. Kawata, C. M. Kortis, R. A. Friesner, *Efficient recursive implementation of the modified Broyden method and the direct inversion in the iterative subspace method: Acceleration of self-consistent calculations*, J. Chem. Phys 108, 11, p. 4426, 1998.

[116] R. A. Kendall, E. Aprà et al., *High Performance Computational Chemistry: An overview of NWChem, a distributed parallel application*, Commun. Comput. Phys. 128, p. 260, 2000.

[117] M. Klein, A. Martinez , R. Seiler, X. P. Wang *On the Born-Oppenheimer Expansion for Polyatomic Molecules* Commun. Math.Phys. 143, p. 607, 1992.

[118] H. M. Klie, *Krylov-secant methods for solving large-scale systems of coupled nonlinear parabolic equations*, PhD thesis, Rice University Houston, TX, USA, 1997.

[119] W. Klopper, F. R. Manby, S. Ten-no, E. F. Vallev, *R12 methods in explicitly correlated molecular structure theory*, Int. Rev. Phys. Chem. 25, p. 427, 2006.

[120] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Russian J. Numer. Anal. Math. Modelling 2, p. 371, 1987.

[121] A. Knyazev, K. Neymeyr, *A geometric theory for preconditioned inverse iteration. III: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl. 358(1-3), p. 95, 2003.

[122] A. Knyazev, K.Neymeyr, *Gradient flow approach to geometric convergence analysis of preconditioned eigensolvers*, SIAM J. Matrix Analysis 31, p. 621, 2009.

[123] O. Koch, C. Lubich, *Dynamical Low Rank Approximation*, SIAM Journal on Matr. Anal. and Appl. 29, 2, p.434 , 2008.

[124] W. Kohn, *Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms*, Phys. Rev. Lett. 76, p. 3168, 1996.

[125] W. Kohn, L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev., 140, p. A113, 1965.

[126] G. Kresse, J. Furthmüller, *Efficient iterative schemes for* ab initio *total-energy calculations using a plane-wave basis set*, Phys. Rev. B 54, 16, p. 11169, 1996.

[127] F. Krüger, R. Schneider, *A Linear Scaling Adaptive Algorithm in DFT*, in preparation.

[128] F. Krüger, *Linear skalierende Methoden in der Hartree-Fock und Dichtefunktionaltheorie*, PhD thesis, TU Berlin, in preparation.

[129] S. A. Kucharsky, R. J. Bartlett, *Fifth-order many-body perturbation theory and its relationship to various coupled-cluster approaches*, Adv. Quantum Chem. 18, p. 281, 1986.

[130] K. N. Kudin, G. E. Scuseria, *Converging self-consistent field equations in quantum chemistry - recent achievements and remaining challenges, ESAIM: Mathematical Modelling and Numerical Analysis*, 41, 2, p. 281, 2007.

[131] H. Kümmel, *Compound pair states in imperfect Fermi gases,* Nucl. Phys. 22 , p. 177, 1961.

[132] H. Kümmel, K. H. Lührmann, J. G. Zabolitzky, *Many-fermion theory in expS- (or coupled cluster) form*, Phys. Reports 36, 1, p. 1, 1978.

[133] W. Kutzelnigg, *Pair Correlation Theories*, Mod. theoretical Chem. 3, p. 129, 1977.

[134] W. Kutzelnigg, *Error analysis and improvement of coupled cluster theory*, Theoretica Chimica Acta 80, p. 349, 1991.

[135] W. Kutzelnigg, *Unconventional aspects of Coupled Cluster theory*, in: Recent Progress in Coupled Cluster Methods, Theory and Applications, Series: Challenges and Advances in Computational Chemistry and Physics 11. To appear 2010.

[136] L. D. Landau, E. M. Lifschitz, *Lehrbuch der Theoretischen Physik, Band III: Quantenmechanik*. Adademie-Verlag, Berlin, 1967.

[137] T. J. Lee, G. E. Scuseria, *Achieving chemical accuracy with Coupled Cluster methods, in Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*, Ed. S. R. Langhof, Kluwer Academic Publishers, Dordrecht, p. 47, 1995.

[138] X.-P. Li, R. W. Nunes and D. Vanderbilt, *Generalization of the density-matrix method to a nonorthogonal basis*, Phys. Rev. B 47, p. 10891, 1993.

[139] E. H. Lieb, B. Simon, *The Hartree-Fock Theory for Coulomb Systems*, Commun. Math. Phys. 53, p. 185, 1977.

[140] E. H. Lieb, *Bound on the maximum negative ionization of atoms and molecules*, Phys. Rev. A 29, 6, p. 3018, 1984.

[141] J. Liesen, P. Tichy, *Convergence analysis of Krylov subspace methods*, GAMM-Mitteilungen 27, 2, p. 153, 2004.

[142] I. Lindgren, J. Morrison, *Atomic Many-body Theory*, Springer, 1986.

[143] P. L. Lions, *Solution of the Hartree Fock equation for Coulomb Systems*, Commum. Math. Phys. 109, 1, p. 33, 1987.

[144] A. Lüchow, J.B. Anderson, *Monte Carlo methods in electronic structure calculations for large systems*, Annu. Rev. Phys. Chem. 51, p. 501, 2000.

[145] S. Lundquist, N. H. March, editors, Theory of the Inhomogeneous Electron Gas, Plenum, 1983.

[146] Y. Maday, G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Num. Math. 94, 4, p. 739, 2003.

[147] J. M. Martinez, T. L. Lopez, *Combination of the sequential secant method and Broyden's method with projected updates*, Computing 25, p. 379, 1980.

[148] D. A. Mazziotti, *Reduced density-matrix mechanics with applications to many-electron atoms and molecules*, Wiley-Interscience, John Wiley & Sons, 2007.

[149] W. Meyer, *PNO-CI studies of electron correlation effects. I. Configuration expansion by means of nonorthogonal orbitals, and application to the ground state and ionized states of methane*, J. Chem. Phys. 58, p. 1017, 1973.

[150] W. Meyer, *PNO-CI and CEPA studies of electronic correlation effects. II: Potential curves and dipole moment functions of the OH radical*, Theoret. Chim. Acta 35, p. 277, 1974.

[151] J. M. Millam, G. E. Scuseria, *Linear scaling conjugate gradient density matrix search as an alternative to diagonalization for first principles electronic structure calculations*, J. Chem. Phys. 106, p. 5569, 1997.

[152] P.M. Morse, H. Feshbach, *Methods of theoretical Physics, Part I: Chapters 1 to 8*, McGraw-Hill Book Company, 1953.

[153] H. Nakatsuji, H. Kato, T. Yonezawa, *On the Unrestricted Hartree-Fock Wavefunction*, Journ. Chem. Phys. 51, 8, p. 3175, 1969.

[154] F. Neese, A. Hansen, D. G. Liakos *Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis*, J. Chem. Phys. 131, 064103, 2009.

[155] J. von Neumann, *Mathematische Grundlagen der Quantenmechanik*, Springer Verlag, Berlin, 1932.

[156] K. Neymeyr, *A geometric theory for preconditioned inverse iteration. I: Extrema of the rayleigh quotient*, Linear Algebra Appl. 332, p. 61, 2001.

[157] K. Neymeyr, *A geometric theory for preconditioned inverse iteration. II: Convergence estimates*, Linear Algebra Appl. 332, p. 87, 2001.

[158] K. Neymeyr,*A geometric theory for preconditioned inverse iteration: IV: On the fastest convergence cases*, Lin. Alg. Appl. 415, 1, p. 114, 2006.

[159] K. Neymeyr, *A geometric theory for preconditioned inverse iteration applied to a subspace*, Math. Comp. 71, p. 197, 2002.

[160] M. Nooijen, K. R. Shamasundar, D. Mukherjee, *Reflections on size-extensivity, size-consistency and generalized extensivity in many-body theory*, Molecular Physics 103, 15-16, p. 2277, 2005.

[161] C. Ochsenfeld and M. Head-Gordon, *A reformulation of the coupled perturbed self-consistent field equations entirely within a local atomic orbital density matrix-based scheme*, Chem. Phys. Lett. 270, p. 399, 1997.

[162] A. J. O'Connor, *Exponential decay of bound state wave functions*, Commun. Math. Phys. 32, p. 319, 1973.

[163] J. M. Ortega, W. C. Rheinboldt, *Iterative Solution of nonlinear equations in several variables*, Acad. Press, 1970.

[164] J. Paldus, *Coupled Cluster Theory*, in: S. Wilson and G.H.F. Diercksen (Eds.), Methods in Computational Molecular Physics, Editors: S. Wilson and G. F. H. Diercksen, Plenum, New York, p. 99, 1992.

[165] R. G. Parr, W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, 1994.

[166] W. Pauli, *The Connection Between Spin and Statistics*, Phys. Rev. 58, p. 716, 1940.

[167] A. Persson, *Bounds for the discrete part of the spectrum of a semibounded Schrödinger operator*, Math. Scand. 8, p. 143, 1960.

[168] P. Piecuch, N. Oliphant, L. Adamowicz, *A state-selective multireference coupled-cluster theory employing the single-reference formalism*, J. Chem. Phys. 99, 1875, 1993.

[169] L. Piela, *Ideas of Quantum Chemistry*, Elsevier Science & Technology, 2007.

[170] P. Pulay, *Convergence acceleration of iterative sequences. The case of SCF iteration*, Chem. Phys. Letters 73, 2, p. 393, 1980.

[171] P. Pulay, *Improved SCF Convergence Acceleration*, Journ. Comp. Chem. 3, 4, p. 556, 1982.

[172] J. Pipek, P. G. Mazay, *A fast intrinsic localization procedure for ab initio and semiempirical linear combination of atomic orbital wave functions*, J. Chem. Phys. 90, 9, p. 4919, 1989.

[173] K. S. Pitzer, *Relativistic effects on chemical properties*, Acc. Chem. Res. 12, 8, p. 271, 1979.

[174] J. A. Pople, R. K. Nesbet, *Self-Consistent Orbitals for Radicals*, J. Chem. Phys. 22, p. 571, 1954.

[175] E. Prugovecki, *Quantum mechanics in Hilbert space*, Second Edition, York/London/Toronto/Sydney/San Francisco, 1981.

[176] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, *A fifth-order perturbation comparison of electronic correlation theories,* Chem. Phys. Lett. 157, p. 479, 1989.

[177] M. Reed, B. Simon, *Methods of Modern Mathematical Physics I - Functional Analysis*, Academic Press, San Diego, 1980.

[178] M. Reed, B. Simon, *Methods of Modern Mathematical Physics II - Fourier Analysis, Self Adjointness*, Academic Press, 1975.

[179] M. Reed, B. Simon, *Methods of Modern Mathematical Physics IV - Analysis of operators*, Academic Press, 1978.

[180] R. Remmert, *Funktionentheorie I*, Springer-Verlag, Grundwissen Mathematik 5, 1984.

[181] T. Rohwedder, R. Schneider, A. Zeiser, *Perturbed preconditioned inverse iteration for operator eigenvalue problems with applications to adaptive wavelet discretization,* Adv. Comp. Math., Springer, to appear; available online, DOI 10.1007/s10444-009-9141-8.

[182] W. Rudin, *Functional Analysis*, Tat McGraw & Hill Publishing Company, New Delhi, 1979.

[183] D. Ruelle, *A remark on bound states in potential scattering theory*, Nuovo Cimento A 61, p. 655, 1969.

[184] R. A. Ryan, *Introduction to Tensor Products of Banach Spaces*, Springer, 2002.

[185] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd edition, SIAM, 2003.

[186] Y. Saad, J. R. Chelikowsky, S. M. Shontz, *Numerical Methods for Electronic Structure Calculations of Materials*, SIAM Review 52, 1, p. 1, 2010.

[187] Y. Saad, M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput. 7 , p. 856, 1986.

[188] B. A. Samokish, *The steepest descent method for an eigenvalue problem with semibounded operators*, Izvestiya Vuzov, Math. 5, p. 105, in Russian, 1958.

[189] U. Scherz, *Quantenmechanik - Eine Einführung mit Anwendungen auf Atome, Moleküle und Festkörper*, Teubner Verlag, 1999.

[190] R. Schneider, *Analysis of the projected coupled cluster method in electronic structure calculation*, Num. Math. 113, 3, p. 433, 2009.

[191] R. Schneider, T. Rohwedder, J. Blauert, A. Neelov, *Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure*, Journal of Comp. Math., 27, p. 360, 2009.

[192] E. Schrödinger, *Quantisierung als Eigenwertproblem*, Vierte Mitteilung, Ann. Phys., Leipzig, 81, p. 220, 1926.

[193] M. Schütz, G. Hetzer, H.-J. Werner, *Low-order scaling local electron correlation methods. I. Linear scaling MP2*, J. Chem. Phys. 111, p. 5691, 1999.

[194] M. Schütz, H.-J. Werner, *Low-order scaling local correlation methods. IV. Linear scaling coupled cluster (LCCSD)*, J. Chem. Phys. 114, p. 661, 2000.

[195] G. E. Scuseria, P. Y. Ayala, *Linear scaling coupled cluster and perturbation theories in the atomic orbital basis*, J. Chem. Phys. 111, p. 8330, 1999.

[196] H. Shull, G. G. Hall, *Atomic units*, Nature 184, p. 1559, 1959.

[197] B. Simon, *Schrödinger operators in the 20th century*, Journal Math. Phys. 41, p. 3523, 2000.

[198] B. Simon, *Schrödinger semigroups*, Trans. Amer. Math. Soc. 7, 1982. Available at `http://www.math.caltech.edu/SimonPapers/R21.pdf` (last visited Feb 4^{th} 2010).

[199] M. H. Stone, *Linear transformations in Hilbert space and their applications to analysis*, Amer. Math. Soc., Colloq. Publ., 1932.

[200] S. Schwinger, *A Posteriori Error Analysis of Effective One-Particle Electronic Structure Calculations*, PhD thesis, MPI Leipzig, in preparation.

[201] A. Szabo, N. S. Ostlund, *Modern Quantum Chemistry*, Dover Publications Inc., 1992.

[202] G. Teschl, *Mathematical Methods in Quantum Mechanics with Applications to Schrödinger Operators*, AMS Graduate Studies in Mathematics 99, 2009.

[203] D. J. Thouless, *Stability conditions and nuclear rotations in the Hartree-Fock theory*, Nuclear Physics 21, p. 225, 1960.

[204] H. A. van der Vorst, C. Vuik, *The superlinear convergence behaviour of GMRES*, Journ. Comp. Appl. Math. 48, 3, p. 327, 1993.

[205] J. Verbeek, J. H. van Lenthe, *The Generalized Slater-Condon Rules, Int. Journ. of quantum chemistry, Vol. XI*, p. 201, John Wiley & Sons Inc., 1991.

[206] J. Weidmann, *Lineare Operatoren in Hilberträumen, Teil I: Grundlagen*, Vieweg u. Teubner, 2000.

[207] J. Weidmann, *Lineare Operatoren in Hilberträumen, Teil II: Anwendungen*, Vieweg u. Teubner, 2003.

[208] F. Wennmohs, F. Neese, *A comparative study of single reference correlation methods of the coupled-pair type*, Chemical Physics 343, p. 217, 2008.

[209] J. Wloka, *Partial differential equations*, Cambridge University Press, reprint, 1992.

[210] T. Yokonuma, *Tensor Spaces and Exterior Algebra*, American Mathematical Society, 1991.

[211] H. Yserentant, *On the electronic Schrödinger equation*. Technical Report 191, SFB 382, Universität Tübingen, 2003.

[212] H. Yserentant, *On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives*, Num. Math. 98, p. 731, 2004.

[213] H. Yserentant, *Sparse grid spaces for the numerical solution of the electronic Schrödinger equation*, Num. Math. 101, p. 381, 2005.

[214] H. Yserentant, *Regularity and Approximability of Electronic Wave Functions*. Book manuscript, to appear in the Lecture Notes in Mathematics series, Springer-Verlag, 2010.

[215] E. Zeidler, *Nonlinear Functional Analysis and Its Applications, Part II B: Nonlinear Monotone Operators*, Springer, 1990.

[216] A. Zeiser, *Direct discretization of the electronic Schrödinger equation on sparse grids*. PhD thesis, TU Berlin, in preparation.

[217] M. Zółkowski, V. Weijo, P. Jørgensen, J. Olsen, *An efficient algorithm for solving nonlinear equations with minimal number of trial vectors: Applications to atomic-orbital based coupled-cluster theory*, J. Chem. Phys. 128, 204105, 2008.